

Handbook OF Research Methods IN Social AND Personality Psychology

Second Edition

EDITED BY

Harry T. Reis

Charles M. Judd

Handbook of Research Methods in Social and Personality Psychology

Second Edition

This indispensable sourcebook covers conceptual and practical issues in research design in the field of social and personality psychology. Key experts address specific methods and areas of research, contributing to a comprehensive overview of contemporary practice. This updated and expanded second edition offers current commentary on social and personality psychology, reflecting the rapid development of this dynamic area of research over the past decade. With the help of this up-to-date text, both seasoned and beginning social psychologists will be able to explore the various tools and methods available to them in their research as they craft experiments and imagine new methodological possibilities.

HARRY T. REIS is Professor of Psychology in the Department of Clinical and Social Sciences, University of Rochester. He is the coauthor of *An Atlas of Interpersonal Situations* and the coeditor of *The Encyclopedia of Human Relationships*.

CHARLES M. JUDD is College Professor of Distinction in the Department of Psychology and Neuroscience at the University of Colorado at Boulder. He is the author of *Data Analysis: A Model Comparison Approach* and *Research Methods in Social Relations*.

Handbook of Research Methods in Social and Personality Psychology

Second Edition

Edited by

Harry T. Reis
University of Rochester

Charles M. Judd
University of Colorado at Boulder



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

32 Avenue of the Americas, New York, NY 10013-2473, USA

Cambridge University Press is part of the University of Cambridge.
It furthers the University's mission by disseminating knowledge in the pursuit of
education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107600751

© Cambridge University Press 2000, 2014

This publication is in copyright. Subject to statutory exception and to the
provisions of relevant collective licensing agreements, no reproduction of any
part may take place without the written permission of Cambridge University
Press.

First published 2000

Reprinted 2009, 2010, 2011

Second edition 2014

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication data

Handbook of research methods in social and personality psychology / [edited by]
Harry T. Reis, University of Rochester,
Charles M. Judd, University of Colorado at Boulder. – Second edition.

pages cm

Includes bibliographical references and indexes.

ISBN 978-1-107-01177-9 (hardback : alk. paper) – ISBN 978-1-10760075-1 (pbk. : alk. paper)

1. Social psychology – Research – Methodology. 2. Personality – Research – Methodology. I. Reis, Harry T.

II. Judd, Charles M.

HM1019.H36 2013

302.07'2–dc23 2013024738

ISBN 978-1-107-01177-9 Hardback

ISBN 978-1-10760075-1 Paperback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party Internet Web sites referred to in this publication and does not guarantee that any content on such Web sites is, or will remain, accurate or appropriate.

Contents

Contributors

Introduction to the Second Edition

Introduction to the First Edition

Harry T. Reis and Charles M. Judd

1 Scratch an Itch with a Brick: Why We Do Research

Susan T. Fiske

Part one. Design and Inference Considerations

2 Research Design and Issues of Validity

Marilynn B. Brewer and William D. Crano

3 Research Design

Eliot R. Smith

4 Causal Inference and Generalization in Field Settings: Experimental and Quasi-Experimental Designs

Stephen G. West, Heining Cham, and Yu Liu

5 Field Research Methods

Elizabeth Levy Paluck and Robert B. Cialdini

Part two. Procedural Possibilities

6 Using Physiological Indexes in Social Psychological Research

Jim Blascovich

7 Research Methods in Social and Affective Neuroscience

Elliot T. Berkman, William A. Cunningham, and Matthew D. Lieberman

8 Behavior Genetic Research Methods: Testing Quasi-Causal Hypotheses Using Multivariate Twin Data

Eric Turkheimer and K. Paige Harden

9 Methods of Small Group Research

Norbert L. Kerr and R. Scott Tindale

10 Inducing and Measuring Emotion and Affect: Tips, Tricks, and Secrets

Karen S. Quigley, Kristen A. Lindquist, and Lisa Feldman Barrett

11 Complex Dynamical Systems in Social and Personality Psychology: Theory, Modeling, and Analysis

Michael J. Richardson, Rick Dale, and Kerry L. Marsh

12 [Implicit Measures in Social and Personality Psychology](#)

Bertram Gawronski and Jan De Houwer

13 [The Mind in the Middle: A Practical Guide to Priming and Automaticity Research](#)

John A. Bargh and Tanya L. Chartrand

14 [Behavioral Observation and Coding](#)

Richard E. Heyman, Michael F. Lorber, J. Mark Eddy, and Tessa V. West

15 [Methods for Studying Everyday Experience in Its Natural Context](#)

Harry T. Reis, Shelly L. Gable, and Michael R. Maniaci

16 [Survey Research](#)

Jon A. Krosnick, Paul J. Lavrakas, and Nuri Kim

17 [Conducting Research on the Internet](#)

Michael R. Maniaci and Ronald D. Rogge

Part three. Data Analytic Strategies

18 [Measurement: Reliability, Construct Validation, and Scale Construction](#)

Oliver P. John and Veronica Benet-Martínez

19 [Exploring Causal and Noncausal Hypotheses in Nonexperimental Data](#)

Leandre R. Fabrigar and Duane T. Wegener

20 [Advanced Psychometrics: Confirmatory Factor Analysis, Item Response Theory, and the Study of Measurement Invariance](#)

Keith F. Widaman and Kevin J. Grimm

21 [Multilevel and Longitudinal Modeling](#)

Alexander M. Schoemann, Mijke Rhemtulla, and Todd D. Little

22 [The Design and Analysis of Data from Dyads and Groups](#)

David A. Kenny and Deborah A. Kashy

23 [Nasty Data: Unruly, Ill-Mannered Observations Can Ruin Your Analysis](#)

Gary H. McClelland

24 [Missing Data Analysis](#)

Gina L. Mazza and Craig K. Enders

25 [Mediation and Moderation](#)

Charles M. Judd, Vincent Y. Yzerbyt, and Dominique Muller

26 [Meta-Analysis of Research in Social and Personality Psychology](#)

Blair T. Johnson and Alice H. Eagly

Author Index

Subject Index

Contributors

John A. Bargh

Yale University

Lisa Feldman Barrett

Northeastern University and Harvard Medical School

Veronica Benet-Martínez

Pompeu Fabra University

Elliot T. Berkman

University of Oregon

Jim Blascovich

University of California

Marilynn B. Brewer

University of New South Wales

Heining Cham

Fordham University

Tanya L. Chartrand

Duke University

Robert B. Cialdini

Arizona State University

William D. Crano

Claremont Graduate University

William A. Cunningham

University of Toronto

Rick Dale

University of California

Jan De Houwer

Ghent University

Alice H. Eagly

Northwestern University

J. Mark Eddy

University of Washington

Craig K. Enders

Arizona State University

Leandre R. Fabrigar

Queen's University

Susan T. Fiske

Princeton University

Shelly L. Gable

University of California

Bertram Gawronski

University of Texas at Austin

Kevin J. Grimm

University of California

K. Paige Harden

University of Texas at Austin

Richard E. Heyman

New York University

Oliver P. John

University of California

Blair T. Johnson

University of Connecticut

Charles M. Judd

University of Colorado at Boulder

Deborah A. Kashy

Michigan State University

David A. Kenny

University of Connecticut

Norbert L. Kerr

Michigan State University

Nuri Kim

Stanford University

Jon A. Krosnick

Stanford University

Paul J. Lavrakas

Northern Arizona University

Matthew D. Lieberman

University of California

Kristen A. Lindquist

University of North Carolina

Todd D. Little

Texas Tech University

Yu Liu

Arizona State University

Michael F. Lorber

New York University

Michael R. Maniaci

University of Rochester

Kerry L. Marsh

University of Connecticut

Gina L. Mazza

Arizona State University

Gary H. McClelland

University of Colorado

Dominique Muller

Pierre Mendes France University at Grenoble, University Institute of France

Elizabeth Levy Paluck

Princeton University

Karen S. Quigley

Northeastern University and Edith Nourse Rogers Memorial (Bedford) VA Hospital

Harry T. Reis

University of Rochester

Mijke Rhemtulla

University of Amsterdam

Michael J. Richardson

University of Cincinnati

Ronald D. Rogge

University of Rochester

Alexander M. Schoemann

East Carolina University

Eliot R. Smith

Indiana University

R. Scott Tindale

Loyola University Chicago

Eric Turkheimer

University of Virginia

Penny S. Visser

University of Chicago

Duane T. Wegener

The Ohio State University

Stephen G. West

Arizona State University

Tessa V. West

New York University

Keith F. Widaman

University of California

Vincent Y. Yzerbyt

Université catholique de Louvain at Louvain-la-Neuve

Introduction to the Second Edition

When we put together the first edition of this *Handbook*, published in 2000, we scarcely could have imagined the pace with which methodological innovation would occur in social and personality psychology. To be sure, we hoped that the field's relentless pursuit of ever-more creative and precise methods would continue – a pursuit that the book was intended to encourage. Our expectation was that a new edition would be needed somewhere in the far distant future. A mere 13 years later, that time has come. Social-personality psychologists have advanced the frontiers of methodology at a far faster rate than we anticipated, so much so that the prior volume of this *Handbook* no longer did justice to the diverse approaches and methods that define the field's cutting-edge research. With these advances in mind, we set out to provide under a single cover a compendium of the most important and influential research methods of contemporary social-personality psychology.

Our goal for this volume is the same as it was for the prior edition: to inform and inspire young researchers to broaden their research practices in order to ask and answer deeper, more finely grained questions about social life. One sometimes hears that methodological innovation provides little more than an incremental gain on what is already known. In our opinion, this view is short-sighted. As Greenwald (2012) observed, the great majority of Nobel Prizes in the sciences have been awarded for methodological advances rather than for theoretical contributions. This, he reasons, is because of the synergy between methodology and theory: Existing theories point to the need for new methods, which then suggest questions that could not have been envisioned, much less investigated, with older methods. In this way, new methods open the door to better understanding of phenomena.

Social-personality psychologists have always been quick to capitalize on new methods and technical innovations to further their exploration of the processes that govern social behavior. Although the field continues to be criticized for overrelying on laboratory experiments conducted with college student samples, we believe that this criticism is short-sighted. As this volume illustrates, social-personality psychologists conduct research using diverse approaches, ranging

from neuroscientific methods to observational coding of live interaction, from implicit assessments to everyday experience studies, and from priming outside of awareness to population-based surveys. Furthermore, the Internet has made possible access to diverse and specialized samples, an opportunity that social-personality psychologists have eagerly embraced. Add to this the sophisticated insights afforded by new or improved statistical innovations such as dyadic data analysis, mediation analysis, and multilevel models, and it is readily apparent that our theories are built on a rich, complex, and mature empirical foundation.

We suspect that our receptivity to innovation is one reason for the growing popularity and influence of social-personality psychology. Membership in the Society for Personality and Social Psychology has more than doubled since 2000, and social-personality psychologists are now often found in schools of business, medicine, and law. The influence of our work extends well beyond the field's traditional borders, so much so that Yang and Chiu (2009), in an analysis of citation patterns in APA journals, identified social psychology as being positioned at the center of the psychological sciences. We believe that this influence is at least partly attributable to our leadership in championing methodological innovation. For example, Baron and Kenny's classic paper on moderation and mediation, published in the *Journal of Personality and Social Psychology* in 1986, is the most cited article of all time in scientific psychology, with more than 34,000 citations at the time of this writing.

Changes in the field's methodology do not occur in a vacuum, of course. Two important developments have been the rapid increase in digital technology and miniaturization, which have led directly to implicit methods, fMRI, and portable devices for recording details of everyday behavior, as well as in the accessibility of the Internet, which has opened the door to a broader pool of research participants. Other kinds of changes have also been influential. For example, the past decade has seen impressive gains in statistical methods. Although many of these methods are computationally complex, they encourage researchers to ask far more intricate and revealing questions than could be asked with *t*-tests, ANOVAs, and correlations. These changes notwithstanding, careful readers will note that our approach to the research process is still grounded in the basics: a concern for internal validity, an appreciation for the complexity of generalizability, and the realization that the most useful and accurate insights will come from programs of research that incorporate multiple, diverse methods.

An easy way to see the rapid pace of methodological innovation in social-personality methods is to compare this edition of the *Handbook* to its

predecessor. The roster of chapters in the current edition represents an extensive revision from the earlier volume. Twelve chapters are entirely new to this volume, discussing topics whose particulars or importance have emerged since publication of the prior volume. These include treatments of field research, implicit methods, methods for social neuroscience and behavior genetics, research on the Internet, methods for studying emotion and dynamical systems, multilevel models, advanced psychometrics, missing data, and mediation and moderation. An additional introductory chapter presents a compelling picture of why we do research. Readers of the first edition will notice that six chapters have been dropped, not because of diminished relevance but rather because there was no way to include them and still have the space necessary to describe newer methods. The remaining chapters have been, in most cases, thoroughly revamped to reflect recent developments in method or application. We believe that the result depicts state-of-the-art methods in social-personality psychology, at least (we feel compelled to point out) for today.

When the two of us entered the field, in the 1970s, a young social-personality psychologist could be considered well trained after taking two courses in statistics and measurement and one in methods. Fortunately, that is no longer the case; methodological training in most graduate programs is far more extensive and continues for the duration of one's career. Although some may see this as a daunting challenge, we prefer to see it as a sign of the health and vigor of our discipline. Social-personality psychologists are dedicated to obtaining the most enlightening, accurate, and useful understanding of the social world in which we live. Taking advantage of methodological innovation to imagine and address newer, more informative questions is the surest way we know to continue the progress of the past few decades. We hope this volume serves as a springboard for the next generation of theoretical advances in social and personality psychology.

Our every expectation is that the methodological advances in the years since the first edition of this volume will only continue to accumulate in the years ahead. We have little doubt that the future promises more appropriate and sophisticated models of data, greater attention to process and mechanisms, increased insights from neuroscientific explorations, greater attention to data from diverse samples and settings, and increased insights in the measurement of automatic responses. And we have no doubt that there are further advances lurking down the road that will come with some surprise. Accordingly, in another dozen years (or perhaps sooner), we suspect it will be time for a third edition of this volume. One certain prediction that we make is that we will not be

the editors of that edition. But we trust that others will realize the excitement of witnessing the methodological vitality of the field by preparing that next edition. Throughout this volume we have loved providing witness to the advances in research methods mentioned earlier in this paragraph.

References

- Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*, 7(2), 99–108.
- Yang, Y. J., & Chiu, C. Y. (2009). Mapping the structure and dynamics of psychological knowledge: Forty years of APA journal citations (1970–2009). *Review of General Psychology*, 13(4), 349–356.

Introduction to the First Edition

Harry T. Reis and Charles M. Judd

It is no accident, we believe, that many of the most influential methodologists in the behavioral sciences happen to identify themselves as social-personality psychologists. Throughout the methodological literature in psychology, citations to Robert Abelson, Donald Campbell, Thomas Cook, Donald Fiske, David Kenny, and Robert Rosenthal, to name just a few, are ubiquitous. The reason we believe that this is not an accident is that social-personality psychologists have set for themselves a particularly challenging methodological task. Their domain of inquiry concerns all of social behavior, from intergroup relations and large-scale social conflict to dyadic interaction and close relationships. They study individual judgments, cognitions, and affects about social phenomena as well as the evolution of social norms and interdependent behaviors at the level of societies. Most recently, entire cultures, and the belief systems associated with them, have become a major area of interest. And, in the tradition of Kurt Lewin, social-personality psychologists are firmly committed to a rigorous empirical approach to whatever they study. They are convinced that a strong and reciprocal relationship between theory and evidence is fundamental to the acquisition of knowledge: that data demand good theories and that theories demand quality data.

As a result, social-personality psychologists have developed and made use of an extensive array of methodological tools. Although the field is sometimes criticized for an overreliance on laboratory experimentation, in fact the diversity of methodological approaches represented in the leading journals is impressive. From surveys to simulations, from laboratory experiments to daily event recordings, from response latency and physiological measures to think-aloud protocols, and from the Internet and palmtop computers to paper-and-pencil reports, the diversity of research designs and procedures, measurement methods, and analytic strategies that social psychologists employ is, in our view, extraordinary.

Our goal in putting together this *Handbook* was to provide a series of state-of-

the-art presentations spanning both traditional and innovative methods that have moved and continue to move the discipline forward. The product, we believe, documents the incredible wealth of methodological tools that social-personality psychologists have at their disposal. Intentionally, we sought to include chapters that might strike some readers as a bit unusual in a book devoted to research methods. Certainly, some of these topics would not have been included in a book of this sort 20, or perhaps even 10, years ago. So, for example, chapters by Hastie and Stasser on simulation, Collins on studying growth and change, McClelland on transformations and outliers, Bargh and Chartrand on cognitive mediation, Reis and Gable on daily experience methods, and Blascovich on psychophysiological measures are a far cry from the traditional chapters on design, measurement, and analysis that one might routinely expect in a research methods textbook. Several statistics chapters are included because we believe that new developments in statistical methodology make it possible to extract valuable insights about social psychological phenomena from data collected with diverse methods in many different settings.

But then, it was not our goal to provide yet another research methods textbook cataloging standard procedures and principles. Many excellent textbooks serving this function are already available. Although this *Handbook* might well be used as a textbook, our goal was more ambitious than teaching the field's traditional core. Rather, we sought to demonstrate and highlight the tremendous methodological richness and innovativeness to be found in social psychological research, and additionally, to provide social-personality psychologists with resources for expanding the methodological diversity employed in their research.

Such innovation is central to the legacy we have inherited from the field's founders. Social-personality psychologists value their reputation as both rigorous and clever methodologists; indeed, among the behavioral sciences, social psychologists are notorious for their exacting methodological standards and for the pinpoint precision with which the fit of evidence to theory is scrutinized. These practices reflect two considerations: the growth of a cumulative literature, which allows researchers to ask ever-finer questions about phenomena and their mediators and moderators, and the availability of new technologies capable of providing information not even imagined a generation or two ago. For example, researchers rarely investigated questions of mediation in the 1960s. With the advent of computerized tests of cognitive mediation, sophisticated measures of physiological mediation, and co-variance structure methods for evaluating mediational models, these questions have become commonplace. A guiding principle in preparing this volume was that theoretical and methodological

questions are not independent. Theory leads us to choose and extend existing methods and search for new tools; methods get us thinking about new ways to test and refine our constructs.

One of Donald Campbell's seminal and lasting contributions is the notion that validity is achieved only through triangulation, by using a variety of methodological approaches and procedures. In its original formulation, this argument primarily addressed the validity and reliability of measurement: through multiple diverse indicators one could eliminate both random and systematic measurement errors and arrive at more accurate appraisals of underlying constructs (e.g., Campbell & Fiske, 1959). We, as researchers, were taught that such a multifaceted measurement approach ought to be employed in each and every study that we conducted.

The discipline is coming to realize that this sort of triangulation is fundamental not simply in measurement but in all aspects of methodology. In this sense, then, it is fitting that the first chapter in this *Handbook*, by one of Donald Campbell's students, Marilynn Brewer, sets the tone for the entire volume. Brewer argues that only through the use of multifaceted research strategies, adopted not only within individual studies but also, and much more important, across an entire program of research, is research validity in its broadest sense achieved. All the diversity that is represented in this volume, and the diversity of methods and approaches yet to be developed, is essential if social-personality research is to produce valid findings, defining validity in its most comprehensive sense: that our conclusions and theories ultimately provide accurate understandings of the social world that we inhabit.

Putting together this volume has inspired in us great pride in social-personality psychology's commitment to methodological rigor and innovation, as well as in the methodological richness of contemporary social psychology. Our hope is that this volume will similarly inspire both new and established researchers alike to broaden and enhance their methodological practices. Additionally, we hope the volume will serve as a stimulus for yet unknown approaches and procedures that further contribute to the validity of the research we conduct. Our legacy as social-personality psychologists mandates that we continue to capitalize on methodological and technological innovations in the service of ever more informative and useful theories and knowledge.

Reference

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.

Chapter one Scratch an Itch with a Brick

Why We Do Research Susan T. Fiske

What do mosquitoes, bricks, and research have in common? Both mosquitoes and research motivate us by bugging us, and both bricks and research build things. But bricks and mosquitoes? Let's see....

Start with the first pair, mosquitoes and research: Both make us itch. Most relevant here, we do research to scratch a mental itch. This is not trivial. Research is challenging; indeed, some would say that personality and social psychology are the really hard sciences, so this handbook provides guidance in doing them right and managing the setbacks. With so much grief (data can be so uncooperative, and reviewers almost always are), you have to have a real itch to do the science, to persist. If research is so tough, we as scientists have to be compelled, have to really want to do it. This chapter explores why and how we bother, brick by brick. So in effect, we are scratching the research itch with a research brick.

When researchers explain how they got involved with particular lifelong projects, they usually answer with some version of, "What really bugged me was this...." Gaps, mysteries, and inconsistencies all drive regular people as much as researchers. Witness the popularity of mystery series, Sudoku puzzles, and suspense genres. People are wired to detect discrepancies and want to resolve them. One prime way to start a program of research is precisely to mind the cognitive gap. That is, scientists especially notice theoretical discrepancies, empirical inconsistencies, missing links in evidence, counterintuitive patterns, and all manner of knowledge that just does not fit (Fiske, 2004a). Noticing discrepancies could be indexed by the still, small buzz at the back of the mind, which interrupts the flow of reading, listening, watching, and synthesizing science. Focusing on the discrepancies is the first step to noticing an unsolved problem. If the discrepancy matters to scientists (for reasons we explore next), they itch to resolve it. And we scratch it by building science, laying the bricks.

This chapter argues that we do research partly to represent our own new perspective on what's missing and what needs to be done. We do this gap-filling

empirically, not just theoretically, because we are a science that does not separate theory and research as much as, for example, theoretical and applied physics or economics. Hence, social and personality scientists mostly do not entertain theoretical contributions without empirical evidence; we are not satisfied until we do the research. As we will see, another separate and not as noble, but very human, motivation for research is that, for those in the field, research is pragmatic in several respects, as people forward their careers. But the most important reasons are intellectual and scientific, so the chapter turns to those first.

Represent New Perspectives

Researchers make discoveries; we create new knowledge. What we bring to our work is our own unique perspective, whether intellectual, personal, identity-based, or even ideological. Some are more conventional sources of science than others, but all form parts of the picture; let's examine each in turn.

Intellectual Puzzles

If science starts with an itch, a discrepancy, or a discontent, we build or use a theory to test explanations. We may detect gaps in existing theory, and this is the platonic ideal for science, as many chapters in this volume illustrate.

Alternatively, researchers may pit two theories against each other, sometimes supporting one to the exclusion of the other, but more often determining the conditions under which each is true. For example, in close relationships research, one might pit attachment theories (Shaver & Mikulincer, 2010) against interdependence theories (Rusbult, Drigotas, & Verette, 1994), but in fact both can operate simultaneously, one at an individual-difference level and the other at a situational level. Still, to the extent that two theories make distinct predictions, the suspense often captures a researcher's (and a reader's) imagination.

Some researchers commit to a meta-perspective, such as evolutionary or functional explanations, and apply them to the problem at hand, building support for that perspective. For example, an evolutionary approach might argue that people mistrust out-groups because it has often been adaptive to stick with your own kind (Neuberg & Cottrell, 2008), and specific research questions follow from these principles.

Another intellectual strategy borrows a neighboring field's theories and

methods, applying them to social and personality phenomena. For example, social cognition research originally began by applying nonsocial models of attention, memory, and inference to social settings, discovering where common principles did and did not apply (see Fiske & Taylor, 2013 for more specific examples). For instance, attention is captured by novel social stimuli, just as by novel nonsocial stimuli (Taylor & Fiske, 1978; McArthur & Post, 1977). However, attention is also captured by information about another's intention (Jones & Davis, 1965), so uniquely social principles sometimes apply to other people, versus things, as objects of perception. So, borrowing from an adjoining field can illuminate what is unique about personality and social approaches.

Still another intellectual strategy of research ideas is going back in time to the earliest psychological writings. Some reread Aristotle (e.g., regarding social animals; Aronson, 2004); some like the French National Archives (e.g., regarding emotion theory; Zajonc, 1985). Myself, I like William James (Fiske, 1992).

Scientists also construct theories from scratch, sometimes going from the top down with a metaphor that seems to capture an important reality, such as depicting willpower as a muscle that can get fatigued (Baumeister & Alquist, 2009). Sometimes theories follow from the bottom up, beginning with data, where a systematic program of research consistently yields particular patterns that demand a systematic explanation. For example, neural responses to face perception suggest that trustworthiness is the first and primary dimension that emerges, and theory then describes why that might be the case (Oosterhof & Todorov, 2008). All these then are intellectual motivators of research.

Personal Experiences

We don't often admit this outside the family, but psychological scientists do often get ideas from personal experience. We are after all part of our own subject matter. Informal sources of formal theory are legitimate, as long as the informal insights are then stated in a systematic and testable form (Fiske, 2004b). Not all theory has to be expressed in mathematical form – indeed, in social and personality psychology, most is not – but it does have to be logical, parsimonious, and falsifiable, unlike common sense. That is, even theory that derives from personal experience has to be accountable to empirical tests.

Being keenly interested in human behavior gives us an advantage in drawing ideas from experience. As trained social observers, we notice behavioral patterns that others miss. Indeed, McGuire (1973) exhorted graduate students to observe

the real, not just what others have said or what the sanitized data say.

Within this approach, the trick is, as Lee Ross puts it, to “run the anecdote” (personal communication, October 12, 2011). If a story, a hunch, or even fiction seems to capture an important human truth, social and personality psychologists can design studies that simulate that phenomenon, to see if it survives the transition from imagination to a reality that replicates reliably. This volume provides instructions for how to do exactly that.

One caveat: New investigators sometimes fall into the trap of doing me-search – that is, studying their own thorny psychological issues, their own in-group's preoccupations, or some intense idiosyncratic experience. The problem here is that, although highly motivated, one may not be the most objective judge of an issue that is too close to home. At worst, one may be too invested in a certain result, and equally bad, one might have no insight at all. At a minimum, the motivational biases we investigate might also bias our interpretation of our results (Kahneman, 2011). At best, one has some relevant insights and an open mind about whether these testable ideas produce interpretable data. Only then is one really ready to learn something scientifically new and reliable, as a result of personal experience.

Group Identities

Many of us go into social psychology because it focuses on the variance explained by situations, and situations can be changed, to benefit people's well-being. If you think a social problem is caused by context, that is potentially a social policy issue, but if you think the social problem has genetic causes, that does not lend itself to easy societal solutions. One important social issue in today's multicultural, globalizing world is intergroup relations – by the author's estimates from conference talk titles, representing the preoccupation of about a quarter to a third of social psychology. As our field itself becomes more heterogeneous, more of us are thinking about various phenomena related to ethnic, racial, cultural, gender, sexual, age, disability, and other diverse identities.

On the principle of “nothing about us without us,” many of the researchers studying these issues come from the affected groups. This presents both opportunities and challenges. The opportunities come in our field's chance finally to represent the underrepresented. Prejudice research, for example, has gone from merely studying the perpetrators to studying the targets, and target-perpetrator interaction (e.g., Richeson & Shelton, 2007), enriching the science of

intergroup interaction, as well as the broader field, with new more widely applicable insights and methods.

The group-identity research faces challenges parallel to the me-search challenges, in what might be viewed as we-search. Besides the perils of lacking objectivity, one is also accountable to a larger identity group, whom one certainly does not wish to alienate with findings that might cast the group in a poor light. This issue arises even more for outsiders studying issues relevant to traditionally oppressed groups, for example, men studying gender and white people studying black experience. Ultimately, membership is not required to conduct good group-related science, but insights do derive from lived experience, and collaboration is one solution to keeping identity-relevant research both sensitive to politics and respectful to lived experience. However, even in these cross-identity collaborations, one must consider whether foregrounding one colleague gains credibility with one audience (e.g., subordinates), and foregrounding the other gains credibility with another audience (e.g., dominants). Peter Glick and I considered this issue in our ambivalent sexism research (e.g., Glick & Fiske, 1996), deciding for this reason, among others (including who ultimately did more work), to foreground the male member of our collaborative team.

Worldview Defense

Even more fraught but also honestly inspiring is research conducted to examine one's own worldview, whether religious, political, or moral. But ideology and science make uncomfortable bedfellows, so this is an enterprise to enter only with both eyes wide open. One has to go into it with the goal of testing cherished assumptions and being willing to find them wanting. For example, liberals and conservatives emphasize distinct moral bases of judgment (Haidt, 2007), and the role of each may unsettle both ends of the spectrum. The inquiry is permissible if one agrees to play by the rules of science. Fortunately, reviewers and editors keep us honest, with no axe-grinding permitted in the ideal case.

Comment

Sources of ideas are as varied as scientists, and we can cluster these sources in various ways. For example, in a classic exhortation to the field, McGuire (1973) listed creative sources as including: paradoxical incident, analogy, hypothetico-deductive method, functional analysis, rules of thumb, conflicting results, accounting for exceptions, and straightening out complex relationships. I do not

disagree, and the interested reader is referred to that earlier account.

Why Run the Study?

All these sources of inspiration are good, but why do research and not just theory? In our field, other scientists will mostly ignore your armchair ideas unless you arrive with evidence in hand. We are trained to be skeptics because ideas are easy; evidence is harder, so it is more precious. What is more, this is science, and when we joined up, we agreed to adhere to the epistemological rule-book. But we also do the studies because research is fun. Let's have a closer look at these motivations to walk the talk, going beyond ideas to research.

Because This Is Science

We do the research because this is science, not theater, law, or car repair. Our rules of evidence appear throughout this volume. When we join a graduate program, we sign on to the scientific norms current in the community of scholars. Reliable evidence that meets shared standards is the coin of the realm.

Social and behavioral sciences might just be, as noted, the truly “hard” sciences, for a variety of reasons. First is measurement: Human reactions are difficult to record because most depend on human observers, whether self-reports on Likert scales or coders of nonverbal behavior, and humans are notoriously unreliable (D. W. Fiske, 1971). As observers of self and other, people are both biased (e.g., prefer to accentuate the positive) and prone to random error (e.g., variable over time, place, modality). Granted, we can use measurements that avoid the human reporter (e.g., reaction time, physiological measures), but these still entail a human judge. Even astronomy recognizes the “personal equation” in observing heavenly bodies (Schaffer, 1988), as apart from human ones. But the celestial stars ultimately submit to more exact measures than the human ones, so finding results in our science – despite the bias, despite the error – is really hard.

Science is all about discovery. Face it, we're geeks; we like making measures, analyzing data, learning stuff. All this is a quest for truth and maybe even wisdom (Brickman, 1980).

Hitting the Sweet Spot Is Fun

The most exciting science finds phenomena of important everyday interest but

connects to old problems for social and personality psychology, which allows well-grounded theory, not just flash-in-the-pan findings popular today but gone tomorrow. Hitting the sweet spot that includes both everyday interest and scientific advance is tricky but fun.

Some advice comes from Stanley Schachter, who reportedly urged his students to craft subtle, seemingly small independent variables that create large, undeniable effects on important dependent variables. For example, handing people a hot rather than iced coffee makes them more generous to strangers (Williams & Bargh, 2008); the warmth variable dates back to early childhood experiences of comfort and safety close to caretakers. What a nifty finding. As another example, making people think about professors makes them better at Trivial Pursuit (Dijksterhuis & van Knippenberg, 1998). This uses everyday materials to make an original point about the power of priming (Bargh & Chartrand, Chapter 13 in this volume).

To hit the sweet spot, another social psychologist, Robert Abelson, counseled young researchers to capture the spirit of the times but before others notice it. Watching trends to get ahead of the curve allows a researcher to anticipate what the field will find interesting next. One does not want to jump on the bandwagon, but rather to drive it, or better yet, to design it. Creating clever, realistic, innovative procedures, which also meet all the criteria of methodological rigor and theory relevance, is indeed fun and motivating. Hitting the sweet spot should make you feel like shouting, “Woo-hoo!”

Solving the Puzzle Is Satisfying

Just as we are bugged by discrepancies and gaps, we like cognitive closure, especially when we have to think a bit to get there. Solving the puzzle is satisfying. Indeed, George Mandler's (1982) theory of aesthetic pleasure posits that people most prefer small discrepancies easily resolved. Musical themes and variations do this. Crossword puzzles do this, if one hits the right level of difficulty. It follows the Goldilocks principle: Not too hard to resolve, not too easy, but moderately challenging seems to work best. One can recognize the right level of difficulty when one notices that time has passed without one being aware of it. Becoming optimally absorbed in the process of puzzle-solving creates the feeling of “flow,” which combines both challenge and skill, resulting in total involvement and complete concentration (Csikszentmihalyi & LeFevre, 1989). This happens more often at work than at leisure, and it makes many of us feel lucky to be paid for what we enjoy most.

Our contributions to the field also are satisfying because they fit previous work, making notable progress, adding to human knowledge, a brick at a time. Both resolving discrepancies and filling the gaps create the “Aha!” experience that keeps problem-solvers going.

Being Right Is Fun

Besides the “woo-hoo” and “aha” experiences, many scientists relish the “gotcha” moment, when they are right about a contested issue. Fun as it is to win a competition, scientists must absolutely fight fair – that is, hurling data, not insults. We all agree to abide by publicly replicable results, although of course interpreting them can remain contentious. In general, in my opinion, picking on other people's results does not usually make the most impact, especially if it is nitpicking. Sometimes, of course, identifying a confounding issue in the established paradigm can release a flood of useful research. Today's methodological side effect can be tomorrow's main effect of interest. This can create cumulative science.

Choosing and framing are essential here. Choose battles carefully: Is the end-result of winning worth making enemies? And if people are challenging your data, try to be a good sport. We are obliged to share our data and any unpublished details of our methods; we must strive to view the challenge as advancing science, to respond vigorously but respectfully to the challengers, who may just improve your work. Keep in mind that they would not be pursuing your findings if they did not consider them important.

If your view ultimately prevails, do not gloat. Apply all the rules of being a good competitive player who respects the other team. These are your colleagues for life, after all. Still, we cannot help rooting for our favored interpretation.

Telling Good Stories Is Entertaining

Good storytellers attract an audience, and our studies are our stories, as witnessed by the popularity of our field with science reporters, best-selling authors, and media moguls. Social and personality psychology can be entertaining, as when our research creates nuggets to share. Although a good science story may sometimes enliven dinner conversation, reciting factoids is probably not a good pick-up strategy. (One might earn the nickname PsycInfo®.) But with a light touch, and presented to the right audience, one might also hear an admiring “Wow!”

Promoting Evidence Is Important

Society needs science. As Daniel Patrick Moynihan reportedly said, “Everyone's entitled to [his] own opinion, not [his] own facts.” Science can inform policy, and if taxpayers foot the bill for our science, we owe them some facts.

What is more, many of us went into the field to try to improve the human condition. We want to identify principles and possibly specific interventions that enhance people's lives. The current federal emphasis on translational research reflects this priority. Our science can improve – or at least inform – social policy. And this too is satisfying.

Sideshows: Pragmatic Reasons for Research

“Is this too idealistic?” you might shrug. We do not just do research because it is exciting, useful, and fun, but also because we have committed to it as a career. Let's acknowledge some practical motivations.

Publish or Perish

We do research partly to get a job. Even if we are hired to teach certain classes, covering certain areas, we are promoted for research published in refereed journals, preferably high-impact ones. Quality, not just quantity, counts here. For example, many tenure, promotion, and award committees consult the h-index (Harzing, 2007), which calculates an author's number of citations relative to the number of total publications, thereby balancing quality and quantity. Journals can also be evaluated this way, to calculate their impact factor, although many journals now use sheer number of downloads, as well as citations, to gauge their status. These indices all tend to converge, which is reassuring for measurement reliability and validity.

Collaborate

Some of the more people-oriented among us do research partly for the rewards of collaboration. When we team up to do science, synergy arrives, joy happens, and companionship shares the inevitable tribulations of the research enterprise. In my humble opinion, cooperation is conducive to good science.

From these teams, we develop networks to connect for friendship and consultation through a career's lifetime. Interdisciplinary collaborative research

in particular often creates the leading edge in science; ideas catch fire when fields rub up against each other, creating the future networks of our sciences. The more social and behavioral scientists learn about the strength of weak ties and the importance of support systems, the more we should seek these linkages in our professional lives. Joint research is one way to do this.

Get Rich (or at Least Get Funding)

Researchers have many intrinsic reasons to seek research funding, not least because it enables them to get their work done. Many schools also emphasize funding as a criterion for promotion because national panels of colleagues have endorsed your research plans. On the pragmatic side, one must have done research to get funding to do research; that is, one must establish a track record. This prior research not only adds credibility, but it also organizes the next steps.

An underappreciated aspect of grant writing is that, even if unfunded, grant proposals organize research. Spending thoughtful effort on a program of research helps one prioritize and manage the ensuing studies, even in the midst of a busy, distracted semester, when the big-picture perspective tends to recede.

Teach

We also do research, among other pragmatic reasons, to inform and motivate our teaching. Contrary to popular belief, teaching and research complement each other. In teaching ratings, research productivity correlates with the professor's rated knowledge, commitment, enthusiasm, and organization (Feldman, 1987). Admittedly, research does not correlate with rated time spent on teaching, there being only so many hours in a day. But students are evidently energized by a teacher who researches.

Serve

Research has unexpected links to service, as well. Our universities want to be famous for the research we do, because quality attracts quality and excitement is contagious, promoting our institutions, who after all write the paychecks. Sometimes we do research to serve populations we cherish (see politics, earlier in the chapter). Sometimes we do research to serve moral causes (also see earlier discussion) or to promote the general health and well-being of humanity.

Be Zen

Researchers rarely consider themselves to be on a spiritual quest, but an often salutary side effect of doing research is being humbled. Data prove us wrong. Students intimidate us with their creativity and hard work. Reviewers undeceive us about the quality of our papers. Peers scoop us and do a better job. The field takes our ideas beyond even our own delusional hopes. Or more often, the field ignores our best ideas (Arkin, 2011). We do not go into research to have these humbling experiences, but they have their own advantages, as just noted.

Most of all, we do research to follow our bliss. We do it because we love it.

Conclusion: Why We Do Research

Research enables us to represent new perspectives, to test our ideas and make them empirically accountable. Along the way, we also discover pragmatic reasons for doing research. The process and our investment in it are knowable and manageable. Besides, we can scratch an itch by laying a brick. Read on, and catch the urge.

References

- Arkin, R. (Ed.). (2011). *Most underappreciated: 50 prominent social psychologists talk about hidden gems* [*Scholarship that missed the mark... misconstruals, misunderstandings, misreporting, misuses, and just plain missed]*. New York: Oxford University Press.
- Aronson, E. (2004). *The social animal* (9th ed.). New York: Worth.
- Baumeister, R. F., & Alquist, J. L. (2009). Self-regulation as a limited resource: Strength model of control and depletion. In J. P. Forgas, R. F. Baumeister, & D. M. Tice (Eds.), *The Sydney symposium of social psychology. Psychology of self-regulation: Cognitive, affective, and motivational processes* (pp. 21–33). New York: Psychology Press.
- Brickman, P. (1980). A social psychology of human concerns. In R. Gilmour & S. Duck (Eds.), *The development of social psychology* (pp. 5–28). New York: Academic Press.
- Csikszentmihalyi, M., & LeFevre, J. (1989). Optimal experience in work and leisure. *Journal of Personality and Social Psychology*, 56(5), 815–822.
- Dijksterhuis, A., & van Knippenberg, A. (1998). The relation between

- perception and behavior, or how to win a game of Trivial Pursuit. *Journal of Personality and Social Psychology*, 74(4), 865–877.
- Feldman, K. A. (1987). Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: A review and exploration. *Research in Higher Education*, 26, 227–298.
- Fiske, D. W. (1971). *Measuring the concepts of personality*. Chicago: Aldine.
- Fiske, S. T. (1992). Thinking is for doing: Portraits of social cognition from daguerreotype to laserphoto. *Journal of Personality and Social Psychology*, 63, 877–889.
- Fiske, S. T. (2004a). Developing a program of research. In C. Sansone, C. Morf, & A. Panter (Eds.), *Handbook of methods in social psychology* (pp. 71–90). Thousand Oaks, CA: Sage.
- Fiske, S. T. (2004b). Mind the gap: In praise of informal sources of formal theory. *Personality and Social Psychology Review*, 8, 132–137.
- Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture* (2nd ed). London: Sage.
- Glick, P., & Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70, 491–512.
- Haidt, J. (2007). The new synthesis in moral psychology. *Science*, 316(5827), 998–1002.
- Harzing, A.W. (2007). *Publish or Perish*, available from <http://www.harzing.com/pop.htm>.
- Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 220–266). New York: Academic Press.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus, & Giroux.
- Mandler, G. (1982). *Mind and emotion*. New York: Krieger.
- McArthur, L. Z., & Post, D. L. (1977). Figural emphasis and person perception. *Journal of Experimental Social Psychology*, 13, 520–535.

- McGuire, W. J. (1973). The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology*, 26(3), 446–456.
- Neuberg, S. L., & Cottrell, C. A. (2008). Managing the threats and opportunities afforded by human sociality. *Group Dynamics: Theory, Research, and Practice*, 12(1), 63–72.
- Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 105(32), 11087–11092.
- Richeson, J. A., & Shelton, J. N. (2007). Negotiating interracial interactions: Costs, consequences, and possibilities. *Current Directions in Psychological Science*, 16(6), 316–320.
- Rusbult, C. E., Drigotas, S. M., & Verette, J. (1994). The investment model: An interdependence analysis of commitment processes and relationship maintenance phenomena. In D. J. Canary & L. Stafford (Eds.), *Communication and relational maintenance* (pp. 115–139). San Diego, CA: Academic Press.
- Schaffer, S. (1988). Astronomers mark time: Discipline and the personal equation. *Science in Context*, 2, 101–131.
- Shaver, P. R., & Mikulincer, M. (2010). New directions in attachment theory and research. *Journal of Social and Personal Relationships*, 27(2), 163–172.
- Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attribution: Top-of-the-head phenomena. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 249–288). New York: Academic Press.
- Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science*, 322(5901), 606–607.
- Zajonc, R. B. (1985). Emotion and facial efference: A theory reclaimed. *Science*, 228(4695), 15–21.

Part one Design and Inference Considerations

Chapter two Research Design and Issues of Validity

Marilynn B. Brewer and William D. Crano Validity refers to “the best available approximation to the truth or falsity of *propositions*” (Cook & Campbell, 1979, p. 37; italics added). In this sense, we cannot speak of the validity or invalidity of research per se. Rather, it is the statements, inferences, or conclusions we wish to draw from the results of empirical research that can be subject to validation. Of course, the way a study is designed and conducted has much to do with the validity of the conclusions that can be drawn from its results, but validity must be evaluated in light of the *purposes* for which the research was undertaken in the first place.

Research Purpose and Types of Validity

The various objectives of research can be classified in any number of ways, but for present purposes the goals of empirical research in social psychology can be differentiated into three broad categories: demonstration, causation, and explanation.

Research performed for the purpose of *demonstration* is conducted to establish empirically the existence of a phenomenon or relationship. Much demonstration research is intended to be descriptive of the state of the world. It includes the frequency of occurrence of specified events across time or space (e.g., distribution of forms of cancer, variations in crime rates, probability of intervention in emergency situations, participation in collective demonstrations, and so forth) and the assessment of the degree of relationship between specified states or conditions (e.g., the correlation between cigarette smoking and lung cancer, the relationship between ambient temperature and violent crime, the correlation between economic prosperity and collective protest, and so on).

Although most descriptive research is conducted in field settings with the purpose of assessing phenomena as they occur “naturally,” some demonstrations also are undertaken in controlled laboratory settings. Studies of gender differences or personality types often are conducted in lab settings. Many of the classic studies in social psychological research – including Sherif's (1935) studies of formation of arbitrary group norms, Asch's (1956) conformity studies, Milgram's (1963) study of obedience to authority, and Tajfel's (1970) initial

studies of ingroup favoritism – were essentially demonstrations of social psychological phenomena conducted in the laboratory.

Although establishing that the presence or absence of one event is correlated with the presence or absence of another is often of interest in its own right, most of the time scientists are interested in whether such covariation reflects a causal relationship between the two events. Thus, much research is undertaken not simply to demonstrate that a relationship exists, but to establish a cause-effect linkage between specific variables (i.e., testing linkages of the form “if X then Y”). For this purpose we are using the concept of causation in the utilitarian sense (Collingwood, [1940](#); Cook & Campbell, [1979](#); Gasking, [1955](#); Mackie, [1974](#)).

From the utilitarian perspective, the purpose of the search for cause-effect relationships is to identify agents that can be controlled or manipulated to bring about changes in outcome. In other words, research on causation is intended to demonstrate that interventions that produce change in one feature of the environment will produce subsequent changes in the outcome of interest. For this purpose, the goal of research is to establish causal connections, not to explain how or why they occur (Cook & Shadish, [1994](#)). For example, in an applied prevention context, Crano, [Siegel, Alvaro, and Patel \(2007\)](#) randomly assigned sixth and seventh grade students to view one of two anti-inhalant-drug messages that were disguised as advertisements accompanying a longer school-based presentation on the dangers of bullying. Results revealed that the participants exposed to an indirectly focused ad evaluated it significantly more positively than did those who received an ad addressed directly to adolescents. This research was designed to test a causal connection, namely whether the change in presentation focus resulted in differences in evaluations of the message, and the experimental design allowed a causal interpretation of the results.

When research has the purpose of establishing causal relationships in this utilitarian sense, the purported causal factor is generally referred to as the independent variable and the outcome or effect as the dependent variable. In fact, the use of these terms in describing a study is effectively a statement of purpose. However, there are important differences across types of research in the meaning of independent variable – differences that have to do with how variation in the purported causal variable is produced. When the state of the independent variable is manipulated by interventions under the control of the researcher, we have research that can be defined as experimental or “quasi-

experimental” (Campbell & Stanley, 1963, 1966; see Chapter 4 in this volume). In correlational field studies, by contrast, the so-called independent variable is not manipulated or controlled; instead, variations are assessed as they occur naturally. Variations are studied for purposes of establishing the relationship between them and subsequent variations in the outcome variable of interest. In such cases, causal inference is usually predicated on temporal precedence, establishing that variations in the purported cause precede variations in the purported effect.

Such temporal precedence is a necessary but not sufficient basis for inferring causation. In studies of this type, the independent variable(s) are more appropriately labeled predictor variables. For example , Hemovich, Lac, and Crano (2011) sought to understand the association of risk variables linking family structure with adolescent substance misuse. In a large-scale secondary-data analysis, youth from dual-parent households were found to be less likely than their single-parent household counterparts to use drugs, and were monitored more closely than single-parent youth. A path analytic model developed to illuminate this relation revealed that family income and structure were associated with parental monitoring, which in turn was linked to adolescents’ social and interpersonal perceptions of drug use, and both of these variables anticipated adolescents’ actual drug use one year later. In this example, temporal precedence and prior theory were used to infer causation, although there was no experimental manipulation of the primary predictor variable (family structure).

As a goal of research, utilitarian causation is sufficient for most applied and action research purposes. Knowing that a reliable cause-effect relationship between X and Y exists is a critical step in designing interventions that can bring about desired changes in the outcome. For utilitarian purposes, what works is what counts, irrespective of why it works. For more basic, theory-testing purposes, knowing that a cause-effect relationship exists is not enough. The purpose of this type of research is *explanation*, or establishing the processes that underlie the linkage between variations in X and Y. This reflects the “essentialist” conceptualization of causation to which most scholars now subscribe (Cook & Campbell, 1979). Research undertaken for the purpose of explanation has the goal of determining not only whether causation exists but how, why, and under what conditions.

Although there are many legitimate questions about validity that can be raised in connection with conclusions drawn from demonstration research (see, e.g., Orne & Holland's [1968] critique of the ecological validity of the Milgram

experimental paradigm), most of the controversies that arise over validity issues in social psychology revolve around inferences about causation and explanation. It was specifically in connection with research intended to establish cause-effect relationships that Campbell introduced the now-classic distinction between internal and external validity (Campbell, 1957; Campbell & Stanley, 1963).

Internal validity, in Campbell's terms, refers to the truth value that can be assigned to the conclusion that a cause-effect relationship between an independent variable and a dependent variable has been established within the context of the particular research setting. The question here is whether changes in the dependent measure were produced by variations in the independent variable (or manipulation, in the case of an experiment) in the sense that the change would not have occurred without that variation. *External validity*, in Campbell's original terminology, referred to the generalizability of the causal finding, that is, whether it could be concluded that the same cause-effect relationship would be obtained across different participants, settings, and methods.

In a later elaboration of validity theory, Cook and Campbell (1979) further differentiated the concept of external validity. They added the term *construct validity* to refer to the extent to which a causal relationship could be generalized from the particular methods and operations of a specific study to the theoretical constructs and processes they were meant to represent. The term “external validity” was reserved to refer to the generalizability of findings to target populations of persons and settings. It is this tripartite distinction – internal, external, and construct validity – that provides the basis for organizing the discussion of validity issues in this chapter.

Internal Validity: The Third-Variable Problem

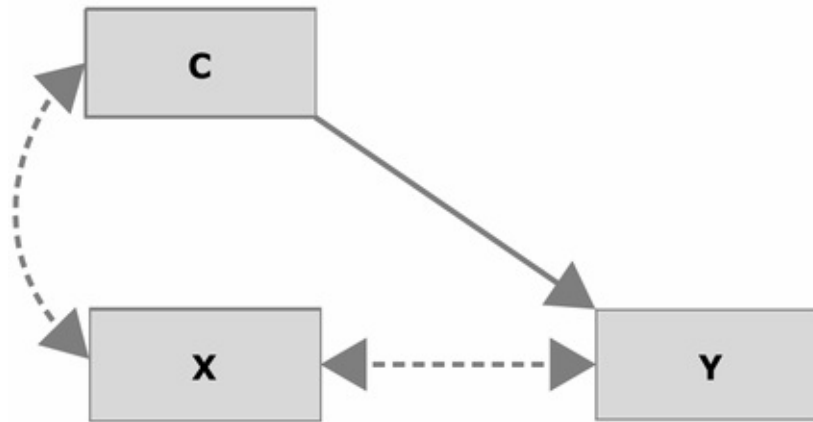
The essence of internal validity is establishing that variation in an effect or outcome (the dependent variable) has been produced by changes in level or intensity of the independent variable and not by some other causal force (or forces).¹ We are interested in the proposition that X causes Y, which in

notational form is expressed as:



Threats to the validity of this proposition come from any *plausible* claim that the

obtained variations in the outcome variable (Y) were actually produced by some third factor (usually unobserved or unmeasured) that happened to be correlated with the variations in the levels of X. Again in notational terms, the alternative



proposition is

In this case, the relationship between X and C (the “hidden” third factor) is not a causal one, and this is indicated by the double-headed arrow. However, because X and C are correlated, causes of the variation in Y could be misattributed to X when they were actually produced by C, as indicated by the single-headed arrow. This pattern is referred to as a *spurious* correlation between X and Y.

This third-variable causation pattern is in part responsible for the well-known dictum that “correlation does not prove causation.” Two variables can be correlated with each other because both are correlates of a third factor, even when there is no direct or indirect causal relationship between the first two. Consider, for example, the inverse relation between marriage and suicide rates in Colonial America. This association might lead to the inference that marriage reduces the likelihood of suicide (in a causal sense). But more careful analysis would raise an alternative explanation. Among wealthy families of the time, marriage was as much a financial match as a romantic one, so when the economy soured, men of formerly rich families had difficulty finding “suitable” mates. A bad economy would depress marriage rates, and we know that suicides spike in bad economic times as well. So, if we correlated marriage and (lagged) suicide rates in Colonial America, we would find a significant inverse relation between these variables. This same relationship could be found across any time period; even so, it certainly cannot be interpreted unambiguously as causal, because of the “hidden” third factor of economics, which is related both to a decrease in marriage rates and an increase in suicides. Thus, the conclusion that marriage acts as a deterrent to suicide would have low internal validity. We could not assign it a high truth value with any confidence.

In social psychological research, many potentially problematic third variables are associated with self-selection. Causal inference is undermined if exposure to different levels of a “treatment” variable is correlated with differences among people in personality or aptitudes. If persons choose for themselves what experiences they will be exposed to, we may observe a relationship between the experience (treatment) and the outcome variable (e.g., persons who engage in intergroup contact are less prejudiced; individuals who travel to hear a Democratic campaigner's speech are more likely to vote Democratic). In these circumstances, we cannot tell whether the outcome was influenced by the treatment or if it would have occurred because of the correlated individual differences even in the absence of the treatment experience.

To establish the causal relationship between two variables unequivocally, variation in the causal factor has to be produced or observed under conditions that are isolated from third factors that may produce a spurious correlation. These third variables must either be held constant or uncorrelated with variations in X. This is the essence of the logic of good experimental design. In addition to control over variation in the independent variable, random assignment of participants to different levels of the manipulated factor serves to rule out many potential third-variable threats to causal inference, particularly self-selection. Without random assignment, manipulating the independent variable is not sufficient to achieve the internal validity of a true experiment (see [Chapter 3](#) of this volume for further discussion of this point). This does not mean that correlational or quasi-experimental studies in field settings can never lead to justified causal inferences (see [Chapter 4](#)). However, many potential threats to the internal validity of such inferences have to be ruled out one by one, whereas random assignment rules out whole classes of potential spurious causal variables in one operation.

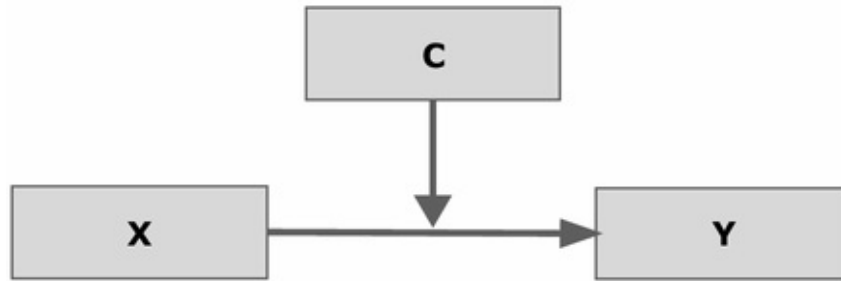
Other Third-Variable Effects: Mediators and Moderators Relationships defined as spurious are ones where there is no true underlying causal linkage between the independent and dependent variables because the true causal variable is some third factor. This should not be confused with other types of causal structures where a third variable is implicated in the relationship between independent and dependent variables. In some cases, a third variable may be part of the causal chain linking X to Y. *Mediated* relations such as these are represented as follows:



In this case, the presence of C is necessary to complete the causal process that links X and Y. In effect, varying X causes variations in C, which in turn cause changes in Y. To return to our example, marriage may indeed have a deterrent effect on suicide, but this effect may be mediated by many factors associated with marriage. Marriage may expose the couple to a wider circle of friends and family, reducing social isolation that may often produce or exacerbate depression. However, being single may not be the only cause of social isolation. In this case, marital status as an independent variable is a sufficient, but not necessary, cause in its link to depression and suicide. To demonstrate that X causes Y only if C occurs does not invalidate the claim that X and Y have a causal relationship; it only explicates the causal chain.

Hidden causes are not the only way that third variables can influence the validity of cause-effect inferences. Sometimes causal relationships can be augmented or blocked by the presence or absence of factors that serve as *moderator variables* (Baron & Kenny, 1986). In social psychological research, for example, attitudes are presumed to be a cause of behavior. We like ice cream (attitude), so we eat ice cream (behavior). However, attitudes sometimes fail to predict related actions. For example, in earlier research, college-age students in Michigan were found to be largely opposed to a proposed law to raise the drinking age from 18 to 21 years (Sivacek & Crano, 1982); however, the relation between their negative attitudes and their willingness to work to defeat the law was weak. Why did their attitudes not impel attitude-consistent actions? The research found the relation between attitudes and actions was affected (moderated) by the extent to which respondents would be personally affected by the law's change. Although college seniors did not like the law, most were not affected by it (i.e., they would be 21 years old before the law came into effect), and so were much less willing to act on their attitude than were those who would be 19 or younger when the law changed. Vested interest, operationalized by age of the respondent, served as a moderator of the attitude-behavior link. These findings do not mean that the attitude-behavior relation is spurious. The moderator variable (age, or vested interest) did not cause the effect (working to defeat the law) in the absence of the independent variable (attitude).

Moderator relationships can be represented notationally as follows:



As shown, the causal link is actually between X and Y, but the observed relationship between these two variables is qualified by levels of variable C, which either enhances or blocks the causal process. Such moderators determine whether the relationship between X and Y that we observe under one set of circumstances (C) would be replicated if those circumstances were changed. Thus, moderators influence *external validity* of any particular research finding, as will be discussed.

Construct Validity: From Construct to Operation and Back Again

For many applied research purposes it is sufficient to know that a specific intervention (e.g., passage of a particular gun control law) produces a specific outcome (e.g., reduction in violent crime). Much social psychological research, however, is inspired not by such specific action-oriented questions but by general theories about the interrelationships among cognition, affect, and social behavior. Theories are stated in terms of abstract concepts and hypothetical constructs that cannot be directly observed or measured. To be subject to empirical testing, theoretical constructs must be “translated” from the abstract to the more concrete, from concepts to operations that can be observed and replicated.

Most social psychological researchers accept the philosophy that the specific operations and measures employed in a given research study are only partial representations of the theoretical constructs of interest – and imperfect representations at that. Hence, the conduct of theory-testing research has a cyclical nature of the form illustrated in [Figure 2.1](#).



Figure 2.1. The cycle of theory-testing research.

The first link in the figure refers to the stage of translating initial theoretical concepts into empirically testable hypotheses and specific operations, and the second link refers to the process of inference from empirical results back to the level of theoretical concepts, which then become the starting point for the next cycle of research.² Construct validity refers to inferences made at both stages of research linking concepts to operations. At the front end, we can ask how valid are the specific operations and measures used in a research project as representations or manifestations of the theoretical constructs to be tested; in other words, how good is the logic of translation from concept to operation? At the last stage, inference goes in the other direction (from empirical operations to hypothetical constructs and processes), and the question becomes how justified is the researcher in drawing conclusions from the concrete findings to the level of theory.

The validity issues that mark the initial, operationalization stage of research are represented in Figure 2.2. Link 1 refers to the inferential link between the operational definition of the independent variable in an experiment and the corresponding causal concept at the theoretical level. Link 2 refers to the analogous link between the hypothetical effect and the actual response measure assessed in an experiment. (A third link could be added to this system to refer to the linkage between assessment of mediating variables and the hypothetical mediational processes, because measures of process are now common in social psychological research.)

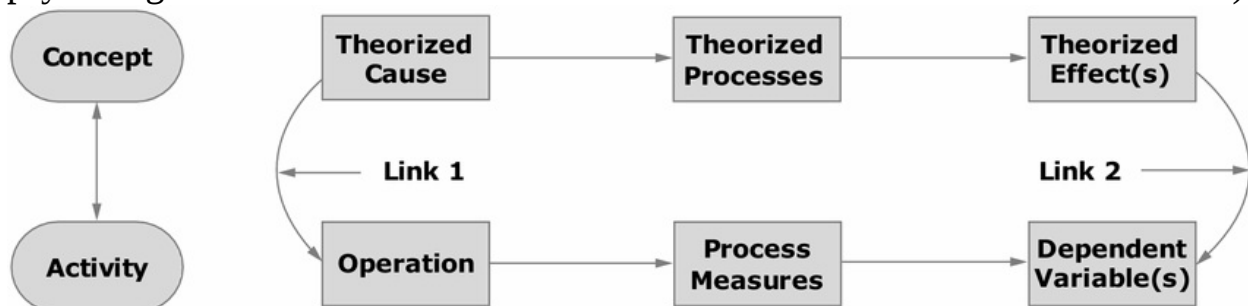


Figure 2.2. Links between theory and operation. Adapted from Rakover (1981).

Rakover (1981) claimed that all of these links are problematic in social psychological research because there are no standardized operations that correspond closely to our hypothetical constructs, and the inferential steps between concept and operation are often quite remote, representing little more than “intuited causal relationships” (p. 125). More specifically, he identified four major difficulties in connecting theory and data: the stimulus and response

validity problems, and the “unknown range of stimulus variation” and “unknown range of measurement” problems.

The stimulus and response validity problems are the standard construct validity questions of whether the stimulus variations and response measures of the empirical research actually reflect variation in the corresponding theoretical states. The unknown-range problems refer to the failure to specify precisely what levels of the independent variable are expected to be causally significant, and over what range of outcomes. Because of these problems, it is difficult to determine whether a failure to confirm a predicted causal or explanatory relationship represents a failure of theory or a failure of operation. The hypothesized relationship could be true at the conceptual level but fail to be demonstrated because operations were unsuccessful or because they failed to fall in the effective range within which the causal process operated.

Causes and Confounds

Criticisms of construct validity often revolve around the meaning of the independent variable as operationalized (Link 1 in [Figure 2.2](#)). Even when the causal efficacy of the independent variable is not in question, there can be questions about the conceptual causal process that is actually operating to produce the observed effect. An independent variable manipulation can involve a host of factors, only one of which may be crucial to the outcome of interest. Typically, such complex variables are successively whittled down with each new study to help isolate the “active ingredient” in the complex intervention, thus allowing greater insight into the critical cause-effect relation. In some applied settings, the question of “Why does it work?” is less important than “Does it work?” In these studies, researchers are content to obtain the sought-for outcome. In more basic, theory-building research, this uncertainty cannot be accepted for long, although it will be tolerated in the initial research phases.

In any research study, the operations meant to represent a particular causal construct can be construed in multiple ways. Any particular operation (manipulation of the independent variable) may be associated with variation in more than one hypothetical state, any one of which may be the “true” causal variable. This is what experimentalists are often referring to when they talk about “confounding” the independent variable: something about the independent variable is causing the outcome of interest, but it is not clear whether it is the construct the researcher had in mind or some other factor that was incorporated in the independent variable.

For instance, a researcher may be interested in the effects of social isolation on susceptibility to influence. An independent variable is designed to produce variations in feelings of social isolation (e.g., waiting in a room with others present, waiting alone for a short time, waiting alone for an extended period of time), but these experimental conditions may also be producing variations in other subjective states, such as fear of the unknown, cognitive rumination, boredom, or fear of interacting with others because of shyness. Causal effects of the “treatment” may be attributable to social deprivation (as intended by the researcher), but also could stem from these other factors that are confounded with isolation in this particular operation.

The causal effect (in the utilitarian sense) of the treatment is not in question in cases such as this, where multiple potential factors are intrinsic to the experimental manipulation, but the *explanation* of the effect is. This type of confounding should be distinguished from threats to internal validity that are not inherent in the independent variable itself. Internal validity is threatened when the independent variable covaries with other variables that are correlated with, but separate from (or extraneous to), the treatment itself. Self-selection, for example, undermines internal validity because individual personality differences have effects that are independent of any effects associated with variations in the intended independent variable. If different types of people select themselves into experimental treatment conditions, then those personality variables that affected their self-selection preferences, rather than the treatment itself, may be responsible for differences between conditions. For example, if an experimental treatment involved the expectation of severe versus mild electric shock, and many subjects refused to participate in the severe but not the mild condition, then differences in outcomes between groups could be caused by the shock variable or by the overrepresentation of brave or foolhardy subjects in the high-shock group. Such potential threats to internal validity can be evaluated or ruled out by examining whether the variations in the independent variable are inadvertently correlated with variations in extraneous variables. Threats to construct validity cannot be so readily disentangled. Nonetheless, there are ways of planning and designing research operations so that the number of potentially confounding factors associated with the independent variable can be reduced.

Many potential confounds arise from the general “reactivity” of social psychological research (Cook & Campbell, 1979) that exists because such research involves social interaction, and subjects usually are aware that they are participants in a research study. Reactivity effects include “demand

characteristics” (Orne, 1962), “experimenter expectancies” (Rosenthal, 1966), and “evaluation apprehension” (Rosenberg, 1969). All of these effects derive from the fact that alert, aware participants are actively seeking cues in the research setting to inform them of what they are expected to do or what they should do to present themselves in a favorable light. Different levels of the independent variable may contain different cues that influence participants’ guesses about what the research study is really about or what constitutes a proper response. When experimental treatments and demand characteristics are confounded in this way, the construct validity of the independent variable is compromised.

We can use the concept of demand characteristics to illustrate the difference between methodological *confounds* (which affect construct validity) and methodological *artifacts* (which are threats to internal validity). Demand characteristics confound the conceptual interpretation of the causal effect of an independent variable when the cues are inherent in the experimental manipulations themselves. To take an example from classic dissonance research, the amount of money participants are offered to write a counterattitudinal essay is intended to manipulate the presence of high or low external justification for engaging in an attitude-discrepant behavior. However, offering a participant as much as \$20 for the favor requested by the experimenter may also carry extraneous cues to the participant that convey the idea that the requested behavior must be either unpleasant or reprehensible to be worth paying so much. In this case, the “message” is implicit in the independent variable itself; \$20 carries a cue or message that is different from an offer of \$5 or \$1. As a consequence, when participants show less attitude change under the high-payment condition than under the low-payment condition, we cannot be sure whether this is owing to the external justification provided by the money offered (the theoretical construct of interest) or to the demand characteristic inherent in the manipulation itself.

Contrast this example with another case in which demand characteristics are created because of experimenter expectancy effects. If a researcher may be biased or predisposed to elicit different responses in different experimental conditions, he or she may deliver the experimental instructions in ways that vary systematically across treatment conditions. For instance, the \$20 offer may be delivered in a different tone of voice or with different nonverbal cues than the \$1 offer. Such experimental behaviors are extraneous to the independent variable itself, but if they are correlated with the differences in instructional conditions, they are procedural artifacts that threaten the internal validity of any causal

interpretations of the effects of the independent variable. This is an illustration of how poor procedural controls can undermine the internal validity of even a true experiment with random assignment to treatment conditions.

Construct Validity and Conceptual Replications

Apart from methodological confounds, research operations are subject to multiple theoretical interpretations. Many interesting controversies in the social psychological literature have been fueled by disagreements over the correct theoretical interpretation of results found in the study of a particular phenomenon. Such debates require conceptual replication, using different operations intended to represent the same causal construct.

Consider, for example, the classic study by Aronson and Mills (1959), in which cognitive dissonance was induced by having female participants read aloud some embarrassing, obscene passages in the guise of an “initiation” test for admission to a discussion group. The intended conceptual independent variable here was a state of dissonance associated with inconsistency between the participant's behavior (going through high embarrassment to join the group) and any negative perceptions of the group. But when participants recite a list of obscene words and then listen to a boring group discussion, one cannot be sure that this experience represents an empirical realization of the intended conceptual variable and nothing else. The complex social situation Aronson and Mills had used has many potential interpretations, including the possibility that reading obscene materials generated a state of sexual arousal that carried over to reactions to the group discussion. If that were the case, it could be that transfer of arousal rather than dissonance accounted for attraction to the group (Zillman, 1983).

A conceptual replication of the initiation experiment by Gerard and Mathewson (1966) was designed to rule out this alternative interpretation. Their experiment differed from the original Aronson and Mills (1959) study in many respects. For example, Gerard and Mathewson used electric shocks instead of the reading of obscene words as their empirical realization of severe initiation (and the dissonance it produced), the shocks were justified as a test of “emotionality” rather than as a test of embarrassment, and the group discussion that participants listened to was about cheating in college rather than sex. Thus sexual arousal was eliminated as a concomitant of the experimental operationalization of the independent variable. The results confirmed the original findings: People who underwent painful electric shocks to become members of a

dull group found that group to be more attractive than did people who underwent mild shocks. Such a confirmation of the basic initiation effect under quite different experimental operations bolstered the contention that it was cognitive dissonance produced by a severe initiation, and not some other conceptual variable, that was responsible for the results in the original experiment. Considerable research in social psychology has been motivated by similar controversies over the valid interpretation of results obtained with complex experimental procedures. Designing conceptual replications to assess threats to construct validity of the causal variable is both challenging and important to the theoretical development of our field.

Multiple Operations: Convergent and Discriminant Validity

These early dissonance experiments illustrate a general principle underlying the idea of construct validity as originally defined by Cook and Campbell (1979), who asserted that the most serious threat to construct validity of any program of research comes from a “mono-operation bias,” that is, the tendency to use only a single operation or measure to represent a particular theoretical construct. Because any one operation invariably underrepresents the construct of interest and embodies potentially irrelevant constructs as well, the conceptual interpretation of single operations can always be challenged. It takes conceptual replication across multiple different operationalizations of the same construct to establish construct validity.

Ideally, multiple operations will allow for testing both convergent and discriminant validity of the construct being studied (Cook & Campbell, 1979). *Convergent* validity is established when different operations representing the same underlying theoretical construct produce essentially the same results, as in the Gerard and Mathewson (1966) experiment described previously. Equally important, however, is establishing that operations that represent the construct of interest show the predicted effect, whereas other operations that do not reflect the theoretical construct do not have similar effects. If Gerard and Mathewson had demonstrated that dissonance aroused by tolerating electric shock had produced attraction to the discussion group, whereas sexual arousal alone (without dissonance) did not produce attraction, they would have gone farther than they actually did in establishing that the dissonance explanation had discriminant validity. That is, dissonance arousal would have been demonstrated to produce effects that differentiated it from other types of arousal.

Measures of dependent variables can also be subjected to the tests of convergent and discriminant validity. To establish measurement construct validity, it is necessary to demonstrate that a particular measure correlates positively with different ways of measuring the same construct and does not correlate as strongly with other measures that use similar methods but are intended to assess a different construct. This is the logic behind the use of the “multitrait multimethod matrix” to establish construct validity of psychological instruments (Campbell & Fiske, 1959). The multitrait-multimethod procedure involves measuring more than one theoretical construct using more than one method for each construct. If a measure has construct validity, different measures of the same trait should be more highly related than different traits assessed by the same method. At a minimum, this logic alerts us to the fact that traits are never measured independent of the method used to instantiate them. At a broader level, the logic can be generalized to testing the theoretical framework within which a construct is embedded. Theoretical validity is established when measured constructs prove to be related to theoretically relevant variables and not to theoretically irrelevant ones, such as the manner in which the variables were measured. Ultimately, then, construct validity is equivalent to theoretical validity.

Causal Processes and Mediation Analyses

Some theoretical debates do not revolve around conceptual interpretation of the operations themselves but are about the intervening processes that mediate the link between the causal variable and its effects. To return to Figure 2.2, these debates over theoretical processes cannot be resolved by examining the construct validity of the independent variable (Link 1) or the dependent variable (Link 2) alone. Theoretical controversies at this level require operations that tap into the intervening physiological, cognitive, and affective processes themselves.

The classic debate between alternative explanations of the counterattitudinal advocacy effect derived from dissonance theory and self-perception theory provides a case in point. In this controversy, the validity of the basic empirical finding and the research operations were not in doubt. Theorists on both sides acknowledged that a causal relationship exists between the presence or absence of external incentives (e.g., monetary payment) and the resulting consistency between behavior and expressed attitudes. At issue was the nature of the mediating processes that underlie the relationship between induced behaviors and subsequent attitudes. Self-perception theorists held that the effect was

mediated by cognitive, self-attribution processes, whereas the dissonance theory explanation rested more on motivational processes. As in many cases in social psychological research, the efforts to establish construct validity helped refine and clarify the theory itself.

Years of attempts to resolve the debate through “critical experiments” were of no avail (Greenwald, 1975). In each case, the same experimental operations could be interpreted as consistent with either theoretical construct. It was not until a clever experiment was designed to assess directly the mediating role of motivational arousal that the deadlock was effectively broken. Zanna and Cooper (1974) used a mediational design to demonstrate that the presence of arousal was necessary to produce the attitude change effects found in the typical dissonance experiment, and that when the motivational effects of arousal were blocked (through misattribution), attitude change following counterattitudinal behavior did not occur. These findings regarding process were more consistent with the dissonance interpretation of the phenomenon than with the self-perception interpretation, although they established that both motivational and cognitive processes were essential mediating factors.

The Many Faces of External Validity

Construct validity represents one form of generalizing from the observed results of an empirical study to conclusions that go beyond the results themselves. Another form of generalizability has to do with the empirical replicability of the phenomenon under study. External validity refers to the question of whether an effect (and its underlying processes) that has been demonstrated in one research setting would be obtained in other settings, at different times, with different research participants, and different research procedures.

In this light, external validity is not a single construct; rather, it is concerned with a whole set of questions about generalizability, each with somewhat different implications for the interpretation and extension of research results. In the sections that follow, we discuss three of the most important forms of external validity: *robustness*, *ecological validity*, and *relevance*. Each raises somewhat different questions about where, when, and to whom the results of a particular research study can be generalized.

Robustness: Can It Be Replicated?

Robustness refers to whether a particular finding is replicable across a variety of

settings, persons, and historical contexts. In its narrowest sense, the question is whether an effect obtained in one laboratory can be exactly replicated in another laboratory with different researchers. More broadly, the question is whether the general effect holds up in the face of wide variations in subject populations and settings. Some findings appear very fragile, obtainable only under highly controlled conditions in a specific context; others hold up despite significant variations in conditions under which they are tested.

Technically, robustness would be demonstrated if a particular research study were conducted with a random sample of participants from a broadly defined population in a random sampling of settings. This approach to external validity implies that the researcher must have theoretically defined the populations and settings to which the effect of interest is to be generalized, and then develop a complete listing of the populations and settings from which a sample is drawn. Such designs are usually impractical and not cost effective. More often, this form of generalizability is established by repeated replications in systematically sampled settings and types of research participants. For instance, a finding initially demonstrated in a social psychology laboratory with college students from an eastern college in the United States may later be replicated with high school students in the Midwest and among members of a community organization on the West Coast. Such replication strategies are practical, and they also have potential advantages for theory testing. If findings do not replicate in systematically selected cases, we sometimes gain clues as to what factors may be important moderators of the effect in question (Petty & Cacioppo, 1996).

Generalizability across multiple populations and settings should be distinguished from generalizability to a particular population. A phenomenon that is robust in the sense that it holds up for the population at large may not be obtained for a specific subpopulation or in a particular context. If the question of generalizability is specific to a particular target population (say, for the elderly), then replication must be undertaken within that population and not through random sampling.

Generalizability from one subject population or research setting to others is one of the most frequently raised issues of external validity for experimental studies conducted in laboratory settings (Henrich, Heine, & Norenzayan, 2010). Sears (1986) provided what probably are the most cogent arguments about the limitations of laboratory experimentation with college student participants. A review of research published in the major social psychology journals in 1985 revealed that 74% of the studies were conducted with undergraduate

participants, and 78% were conducted in a laboratory setting. According to Sears, this restriction of populations and settings means that social psychology has a “narrow data base” on which to draw conclusions about human nature and social behavior.

It is important to point out here that Sears (1986) was not claiming that college students or psychology laboratories are any less generalizable to the world at large than any other specific type of persons or settings. Just because an effect has been demonstrated in a particular field setting rather than a lab does not automatically render it more externally valid, and college sophomores might not differ in critical ways from members of the population at large. Sears was criticizing the overrepresentation of a specific type of participant and setting across many studies, all of which share the same limitations on external validity. However, before we can conclude that the oversampling of college student participants actually limits the external validity of our findings and interpretations, we have to specify in what ways undergraduate students differ from other populations and how these differences might alter the effects we observe.

Drawing on research in cognitive and social development, Sears suggested that there are several distinguishing characteristics of college students that may be relevant to social psychological findings. Compared with the general population, undergraduates are likely to have stronger cognitive skills, less well-formulated or crystallized attitudes and self-concepts, and less stable group identities – differences whose effects are likely to be exacerbated when studies are conducted in academic laboratories with academic-like tasks. Do these differences make a difference? Sears contended that our research may exaggerate the magnitude of effects of situational influences and cognitive processes on social attitudes and behavior because of the characteristics of our subject population and setting in which they typically are tested.

To argue that characteristics of the setting or subject population qualify the conclusions that can be drawn about cause-effect relationships is, in effect, to hypothesize that the cause interacts with (i.e., is moderated by) the characteristics of the population or context to produce the effect in question. To translate Sears's (1986) arguments into these terms, he is postulating that some of the common manipulations used in our social research laboratories may interact with participant characteristics to determine research outcomes. For instance, the magnitude of the effect of influence attempts that rely on cognitive elaboration would be expected to differ depending on whether the effect is tested

with college students or with older, nonstudent populations. In this case, age is expected to moderate the causal effect of treatments that require cognitive elaboration.

More recently, the generalizability of much social psychological research has also been criticized because of the cultural homogeneity of our research participants (Markus & Kitayama, 1991; Sampson, 1977). Historically, the vast majority of experimental social psychology has been conducted in North American and European countries with participants socialized in Western cultural traditions and values. So the question arises as to whether the causal relationships we have identified reflect universal social psychological processes or are culture-specific. As with criticisms regarding age and education of participants, making the case that culture is a threat to external validity of findings requires explicating how cultural differences might influence the causal processes under investigation and then testing whether cultural background of participants moderates (i.e., interacts with) experimental manipulations.

External validity is related to settings as well as to participant populations. The external validity of a finding is challenged if the relationship between independent and dependent variables is altered when essentially the same research procedures are conducted in a different laboratory or field setting or under the influence of different experimenter characteristics. For example, Milgram's (1963) initial studies of obedience were conducted in a research laboratory at Yale University, but used participants recruited from the community of New Haven. Even though these experiments were conducted with a nonstudent sample, a legitimate question is the extent to which his findings would generalize to other settings. Because participants were drawn from outside the university and because many had no previous experience with college, Milgram worried that the prestige and respect associated with a research laboratory at Yale may have made the participants more susceptible to the demands for compliance that the experiment entailed than they would have been in other settings.

To address this issue, Milgram replicated his experiment in a very different physical setting. By moving the research operation to a “seedy” office in the industrial town of Bridgeport, Connecticut and adopting a fictitious identity as a psychological research firm, Milgram hoped to minimize the reputational factors inherent in the Yale University setting. In comparison with data obtained in the original study, the Bridgeport replication resulted in slightly lower but still dramatic rates of compliance to the experimenter. Thus, setting could be

identified as a contributing but not crucial factor to the basic findings of the research.

Cook and Campbell (1979) made it clear that questions of external validity, or generalizability, are implicitly questions about interactions between the independent variable (treatment) and contextual variables such as participant selection, history, and research setting. In other words, the quest for external validity is essentially a search for moderators that limit or qualify the cause-effect relationship under investigation. As the Milgram experiments illustrate, once one has identified what the potential moderators are, the robustness of an effect can be tested empirically by varying those factors systematically and determining whether or not the effect is altered.

Ecological Validity: Is It Representative?

The question of whether an effect holds up across a wide variety of people or settings is somewhat different from asking whether the effect is representative of what happens in everyday life. This is the essence of *ecological validity* – whether an effect has been demonstrated to occur under conditions that are typical for the population at large. The concept of ecological validity derives from Brunswik's (1956) advocacy of “representative design,” in which research is conducted with probabilistic samplings of participants and situations.

Representativeness is not the same as robustness. Generalizability in the robustness sense asks whether an effect can occur across different settings and people; ecological validity asks whether it does occur in the world as is. In Brunswik's sense, findings obtained with atypical populations (e.g., college students) in atypical settings (e.g., the laboratory) never have ecological validity until they are demonstrated to occur naturally in more representative circumstances.

Many researchers (e.g., Berkowitz & Donnerstein, 1982; Mook, 1983; Petty & Cacioppo, 1996) take issue with the idea that the purpose of most research is to demonstrate that events actually *do* occur in a particular population. Testing a causal hypothesis requires demonstrating only that manipulating the cause *can* alter the effect. Even most applied researchers are more interested in questions of what interventions could change outcomes rather than what happens under existing conditions. Thus, for most social psychologists, ecological validity is too restrictive a conceptualization of generalizability for research that is designed to test causal hypotheses. For causal inferences to be ecologically valid, the setting in which a causal principle is demonstrated need not physically

resemble the settings in which that principle operates in real life. As Aronson, Wilson, and Brewer (1998) put it, most social psychology researchers are aiming for “psychological realism,” rather than “mundane realism,” in their experiments. *Mundane realism* refers to the extent to which the research setting and operations resemble events in normal, everyday life. *Psychological realism* is the extent to which the psychological processes that occur in an experiment are the same as the psychological processes that occur in everyday life. An experimental setting may have little mundane realism but still capture processes that are highly representative of those that underlie events in the real world.

Relevance: Does It Matter?

In a sense, the question of ecological validity is also a question of relevance – is the finding related to events or phenomena that actually occur in the real world? However, relevance also has a broader meaning of whether findings are potentially useful or applicable to solving problems or improving quality of life. Relevance in this latter sense does not necessarily depend on the physical resemblance between the research setting in which an effect is demonstrated and the setting in which it is ultimately applied. Perceptual research on eye-hand coordination conducted in tightly controlled, artificial laboratory settings has proved valuable to the design of instrument panels in airplanes even though the laboratory did not look anything like a cockpit. Misunderstanding this principle is the cause of periodic, ill-conceived, politically inspired critiques of “silly” research.

Relevance is the ultimate form of generalization, and differences among research studies in attention to relevance are primarily matters of degree rather than of kind. All social psychological research is motivated ultimately by a desire to understand real and meaningful social behavior. But the connections between basic research findings and applications often are indirect and cumulative rather than immediate. Relevance is a matter of social process, that is, the process of how research results are transmitted and used rather than what the research results are (Brewer, 1997).

Is External Validity Important?

External validity, like other validity issues, must be evaluated with respect to the purposes for which research is being conducted. When the research agenda is essentially descriptive, ecological validity may be essential. When the purpose is utilitarian, robustness of an effect is particularly critical. The fragility and

nongeneralizability of a finding may be a fatal flaw if one's goal is to design an intervention to solve some applied problem. On the other hand, it may not be so critical if the research purpose is testing explanatory theory, in which case construct validity is more important than other forms of external validity.

In the field of physics, for example, many phenomena can only be demonstrated empirically in a vacuum or with the aid of supercolliders. Nonetheless, the findings from these methods are often considered extremely important for understanding basic principles and ultimate applications of the science. Mook (1983) argued compellingly that the importance of external validity has been exaggerated in the psychological sciences. Most experimental research, he contended, is not intended to generalize directly from the artificial setting of the laboratory to “real life,” but to test predictions based on theory. He drew an important distinction between “generality of findings” and “generality of conclusions” and held that the latter purpose does not require that the conditions of testing resemble those of real life. It is the understanding of the processes themselves, not the specific finding that has external validity.

In effect, Mook (1983) argued that construct validity is more important than various forms of external validity when we are conducting theory-testing research. Nonetheless, the need for conceptual replication to establish construct validity requires a degree of robustness across research operations and settings that is very similar to the requirements for establishing external validity. The kind of systematic, programmatic research that accompanies the search for external validity inevitably contributes to the refinement and elaboration of theory as well.

Optimizing Types of Validity

Among research methodologists, controversies have ensued about the relative importance of different validity concerns, ever since Campbell and Stanley (1963) took the position that internal validity is the sine qua non of experimental research and takes precedence over questions of external validity (e.g., Cook & Shadish, 1994; Cronbach, 1982). The debate includes discussions of whether there are necessary trade-offs among the various aspects of validity or whether it is possible to demand that research maximize internal, external, and construct validity simultaneously.

It is possible to conduct a single research study with the goal of maximizing internal validity. Questions of external validity and construct validity, however,

can rarely be addressed within the context of a single research design and require systematic, programmatic studies that address a particular question across different participants, operations, and research settings. Thus, it is patently unfair to expect that any single piece of research have high internal, external, and construct validity all at the same time. It is more appropriate to require that *programs* of research be designed in a way that addresses all types of validity issues.

When designing such a research program, it is important to recognize the ways in which efforts to maximize one form of validity may reduce or jeopardize other types, hence the need for a diversity of methods as represented throughout this volume. By understanding such trade-offs, we can plan research projects in which the strengths and weaknesses of different studies are complementary. Let us consider some of these important complementarities.

Setting: Lab versus Field

It is a common assumption that laboratory research achieves high internal validity at the expense of external validity, whereas research conducted in natural field settings is associated with greater external validity at the cost of more threats to internal validity. There is some basis for this implied association between research setting and types of validity. The laboratory often permits a degree of control of the causal variable that maximizes internal validity to a degree that is difficult to achieve in “noisy” real-world contexts. And to the extent that natural settings reduce the reactivity characteristic of laboratory-based research, one threat to external validity is reduced.

We hope, however, that the earlier discussion of the different types of validity has made it clear that there is no invariable association between the setting in which research is conducted and its degree of internal, external, or construct validity. Tightly controlled experimental interventions can be introduced in field settings (and, conversely, laboratory studies can be poorly controlled). Conducting research in a naturalistic context does not by itself confer external validity, just as conducting an experiment in the laboratory does not itself confer internal validity. Any specific context has limited generalizability. Even if the setting has been chosen to be highly representative or typical of naturally occurring situations, ecological validity is suspect if the research introduces conditions in the setting that do not occur spontaneously.

Establishing either construct validity or external validity requires that the conclusions drawn from research hold up across variation in context. Thus, it is

the complementarity of field and lab as research settings that contributes to validity, not the characteristics of either setting alone. One good illustration of the use of selected field sites in conjunction with laboratory research comes from the literature on mood and altruism. A variety of mood induction manipulations have been developed in laboratory settings, such as having participants read affectively positive or negative passages, watch stimulating movie passages, or think about particularly happy or unhappy events in their lives. After the mood state induction, participants are given an opportunity to exhibit generosity by donating money or helping an experimental accomplice. Results from these lab studies showed that positive mood induction elevates helping behavior whereas depressed mood inhibits helping (e.g., Aderman, 1972).

Despite multiple replications of this effect in different laboratories with different investigators, the validity of these findings has been challenged both because of the artificiality of the setting in which altruism is assessed and because of the potential demand characteristics associated with the rather unusual mood induction experience. To counter these criticisms, researchers in the area took advantage of a natural mood induction situation based on the emotional impact of selected motion pictures (Underwood et al., 1977).

After some pilot research, in which ratings were obtained from moviegoers, a double feature consisting of *Lady Sings the Blues* and *The Sterile Cuckoo* was selected for its negative-affect-inducing qualities, and two other double features were selected to serve as neutral (mildly positive) control conditions. A commonly occurring event – solicitation of donations to a nationally known charity with collection boxes set up outside the movie theater lobby – was chosen as the vehicle for a measure of the dependent variable of generosity.

Having located such naturally occurring variants of the laboratory mood induction operation and altruism measure, the major design problem encountered by the researchers was that of participant self-selection to the alternative movie conditions. Although random assignment of volunteer moviegoers was a logical possibility, the procedures involved in utilizing that strategy would have created many of the elements of artificiality and reactivity that the field setting was selected to avoid. Therefore, the investigators decided to live with the potential problem of self-selection and to alter the research design to take its effect into consideration. For this purpose, the timing of collection of donations to charity at the various theaters was randomly alternated across different nights so that it would occur either while most people were entering the theater (before seeing the movies) or leaving (after seeing both features). The rate of donations given

by arriving moviegoers could then be a check on preexisting differences between the two populations apart from the mood induction. Fortunately, there proved to be no differences in initial donation rates as a function of type of movie, whereas post-movie donations differed significantly in the direction of lowered contribution rates following the sad movies. This pattern of results, then, preserved the logic of random assignment (initial equivalence between experimental conditions) despite the considerable deviation from ideal procedures for participant assignment.

Two points should be emphasized with respect to this illustration of field research. First, the field version of the basic research paradigm was not – and could not be – simply a “transplanted” replication of the laboratory operations. The researchers had considerably less control in the field setting. They could not control the implementation of the stimulus conditions or extraneous sources of variation. On any one night a host of irrelevant events may have occurred during the course of the movies (e.g., a breakdown of projectors or a disturbance in the audience) that could have interfered with the mood manipulation. The researchers were not only helpless to prevent such events but might not have even been aware of them if they did occur. In addition, the field experimenters could not assign participants randomly to conditions and had to rely on luck to establish initial equivalence between groups.

Second, the results of the field experiment as a single isolated study would have been difficult to interpret without the context of conceptually related laboratory experiments. This difficulty is partly attributable to the ambiguities introduced by the alterations in design and partly to the constraints on measurement inherent in the field situation, where manipulation checks, for example, are not possible. The convergence of results in the two settings greatly enhances our confidence in the findings from both sets of operations.

Isolation versus Construct Validity

Laboratory experiments are inherently artificial in the sense that causal variables are isolated from their normal contextual variation. This isolation and control is the essence of testing causal hypotheses with a high degree of internal validity. Isolation does not necessarily jeopardize external validity if the experimental situation has psychological realism, that is, if the causal processes being represented in the lab setting are the same as those that operate in nonlaboratory contexts.

It is this matter of whether the process is the “same” when the context is

altered that constitutes the stickiest issue of validity. Greenwood (1982) called this the problem of “the artificiality of alteration,” the problem that arises whenever bringing a variable into the laboratory changes its nature. Greenwood argued that this alteration is particularly problematic for social psychology because social psychological phenomena are inherently relational or context-dependent and hence do not retain their identity when isolated from other psychological processes. He takes this as a fatal criticism of laboratory research methods in social psychology, but the truth is that it applies to any context in which a phenomenon is observed, as psychological experiences are never exactly the same from one time or place to another.

The issue here is one of the level of abstraction at which constructs or principles are defined. Consider, for example, the construct of “threat to self-esteem.” Most of us would guess that being informed that one has failed a test of creative problem solving would have more impact on self-esteem of a Harvard undergraduate than it would on a 50-year-old mineworker. Thus, if we were interested in the effects of lowered self-esteem on aggression, we might have to use different techniques to lower self-esteem in the two populations. Threats to self-esteem based on challenges to one's academic self-concept are certainly different in many ways from challenges that threaten one's sense of group belonging or physical stamina. But if each of these, in their appropriate context, proves to have an impact on anger or aggressiveness, then we have gained confidence in a general principle that threats to components of self-esteem that are important or central to one's sense of identity increase aggression.

This, then, is the ultimate challenge for valid theory building in social psychology. Our theoretical constructs must be abstract enough to generalize across a range of contexts and specific manifestations, yet precise enough to permit testing at an empirical level. Each empirical demonstration is inevitably limited to a specific context and participant sample, and subject to multiple interpretations. But each time a theoretical proposition is tested in a new setting or with new operations and more diverse participant samples, a contribution is made to the overall picture. Validity is never the achievement of a single research project, but rather is the product of cumulative theory testing and application.

Some Final Points.

Throughout this chapter we have attempted to make some basic points about the multiple meanings of validity with respect to the conclusions drawn from the

results of research studies. Further, we have emphasized that validity has to be evaluated in light of the *purposes* for which the research has been conducted and the kind of inferences that are to be made. Research undertaken to accurately describe the state of the world at a particular time and place must be judged against standards of validity different from research conducted to test a causal hypothesis or an explanatory theory. In this chapter we have focused particularly on research designed to test and explain causal relationships because that is representative of a large proportion of the research conducted and published in social psychology. We have also simplified matters by focusing on the causal relationship between one independent variable (X) and one dependent variable (Y) and have drawn many of our research illustrations from classical studies in social psychology where such single causal hypotheses were under investigation.

By focusing on the simplest two-variable case, we hope we have been able to explicate the conceptual distinctions among the different forms of validity in their clearest form. Admittedly, however, this simple case is not representative of much current research in social psychology, which more often involves multiple independent variables and their interactive effects on multiple dependent variables in the context of complex multivariate causal models. Research designs are also further complicated by looking at relationships over time and across multiple levels of analysis. Each of these expansions of the basic $X \rightarrow Y$ paradigm introduces additional methodological and statistical issues, many of which are covered in detail in other chapters in this volume. But the increased complexity of research designs does not change the fundamental questions of validity that must be addressed in evaluating inferences that are drawn from empirical findings. The distinctions among internal validity, construct validity, and external validity as different criteria for evaluation are just as applicable to conclusions drawn from multivariate, multilevel research as they are for interpreting the simplest bivariate relationship. The same principles of validation apply, but the operations involved in their application become more complex. We believe that these added costs are well worth the effort, and must be borne if we are to make progress in understanding the complex human behaviors that are the very essence of our discipline.

References

- Aderman, D. (1972). Elation, depression, and helping behavior. *Journal of Personality and Social Psychology*, 24, 91–101.
- Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a

- group. *Journal of Abnormal and Social Psychology*, 59, 177–181.
- Aronson, E., Wilson, T., & Brewer, M. B. (1998). Experimentation in social psychology. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 99–142). Boston: McGraw-Hill.
- Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs*, 70 (9, Whole No. 416).
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist*, 37, 245–257.
- Brewer, M. B. (1997). The social psychology of intergroup relations: Can research inform practice? *Journal of Social Issues*, 53(1), 197–211.
- Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: University of California Press.
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297–312.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand-McNally.
- Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.
- Collingwood, R. G. (1940). *An essay on metaphysics*. Oxford: Clarendon.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.
- Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments

- over the past fifteen years. *Annual Review of Psychology*, 45, 545–580.
- Crano, W. D., Siegel, J. T., Alvaro, E. M., & Patel, N. (2007). Overcoming resistance to anti-inhalant appeals. *Psychology of Addictive Behaviors*, 21, 516–524.
- Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.
- Gasking, D. (1955). Causation and recipes. *Mind*, 64, 479–487.
- Gerard, H. B., & Mathewson, G. C. (1966). The effects of severity of initiation on liking for a group: A replication. *Journal of Experimental Social Psychology*, 2, 278–287.
- Greenwald, A. G. (1975). On the inconclusiveness of “crucial” cognitive tests of dissonance versus self-perception theories. *Journal of Experimental Social Psychology*, 11, 490–499.
- Greenwood, J. D. (1982). On the relation between laboratory experiments and social behaviour: Causal explanation and generalization. *Journal for the Theory of Social Behaviour*, 12, 225–250.
- Hemovich, V., Lac, A., & Crano, W. D. (2011). Understanding early-onset drug and alcohol outcomes among youth: The role of family structure, social factors, and interpersonal perceptions of youth. *Psychology, Health & Medicine*, 16, 249–267.
- Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61–83.
- Mackie, J. L. (1974). *The cement of the universe*. Oxford: Oxford University Press.
- Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98, 224–253.
- Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371–378.
- Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, 38, 379–387.
- Orne, M. (1962). On the social psychology of the psychological experiment.

American Psychologist, 17, 776–783.

Orne, M., & Holland, C. (1968). Some conditions of obedience and disobedience to authority: On the ecological validity of laboratory deceptions. *International Journal of Psychiatry*, 6, 282–293.

Petty, R. E., & Cacioppo, J. T. (1996). Addressing disturbing and disturbed consumer behavior: Is it necessary to change the way we conduct behavioral science? *Journal of Marketing Research*, 33, 1–8.

Rakover, S. S. (1981). Social psychology theory and falsification. *Personality and Social Psychology Bulletin*, 7, 123–130.

Rosenberg, M. J. (1969). The conditions and consequences of evaluation apprehension. In R. Rosenthal & R. Rosnow (Eds.), *Artifact in behavioral research* (pp. 279–349). New York: Academic Press.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.

Sampson, E. E. (1977). Psychology and the American ideal. *Journal of Personality and Social Psychology*, 35, 767–782.

Sears, D. O. (1986). College sophomores in the laboratory: Influence of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51, 515–530.

Sherif, M. (1935). A study of some social factors in perception. *Archives of Psychology*, 27(187), 1–60.

Sivacek, J., & Crano, W. D. (1982). Vested interest as a moderator of attitude behavior consistency. *Journal of Personality and Social Psychology*, 43, 210–221.

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American*, 223(2), 96–102.

Underwood, B., Berenson, J., Berenson, R., Cheng, K., Wilson, D., Kulik, J., Moore, B., & Wenzel, G. (1977). Attention, negative affect, and altruism: An ecological validation. *Personality and Social Psychology Bulletin*, 3, 54–58.

Zanna, M., & Cooper, J. (1974). Dissonance and the pill: An attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology*, 29, 703–709.

Zillman, D. (1983). Transfer of excitation in emotional behavior. In J. T. Cacioppo & R. E. Petty (Eds.), *Social psychophysiology: A sourcebook* (pp. 215–240). New York: Guilford Press.

¹ This does not mean that the independent variable under investigation is assumed to be the only cause of the outcome, but rather that this variable has a causal influence independent of any other causal forces.

² Because abstract definitions and theory are rarely unaffected by the process and outcomes of empirical research, we assume here that Construct₁ and Construct₂ are not necessarily conceptually equivalent.

Chapter three Research Design

Eliot R. Smith^{*}

Research design is the systematic planning of research to permit valid conclusions. Design involves, for example, the specification of the population to be studied, the treatments to be administered, and the dependent variables to be measured – all guided by the theoretical conceptions underlying the research. Research design most fundamentally affects the internal validity of research, that is, the ability to draw sound conclusions about what actually causes any observed differences in a dependent measure. However, design also has implications for other forms of validity (Cook & Campbell, 1979). Statistical conclusion validity (especially statistical power) is affected by such design-related issues as the number of participants used in a study and the way they are allocated to conditions. Construct validity – the ability to link research operationalization to their intended theoretical constructs – is affected by many aspects of design, such as freedom from confounding. External validity or generalizability is affected by the way other design factors beside those of key theoretical interest are held constant or allowed to vary. This chapter points out the implications of design for all of these forms of validity, as they become relevant. See Brewer and Crano (Chapter 2 in this volume) for more on the types of validity and their interrelationships.

Focus of This Chapter

Research design is inextricably linked to data analysis. An appropriate design can ensure that the substantive and statistical assumptions for the data analysis, such as the assumptions that permit strong causal inferences, are met. This chapter focuses on design rather than analysis, but will make brief references to important analytic issues. The presentation here is largely nonmathematical. Treatments of the mathematical principles underlying design can be found in Kirk (1968), Keppel (1982), Winer (1971), and other sources. This chapter follows the lead of Abelson (1995), Campbell and Stanley (1963), and Shadish, Cook, and Campbell (2002) in emphasizing logic rather than equations. Also, this chapter downplays the “nuts and bolts” aspects of research procedures, such

as how to construct a plausible cover story and how to administer a manipulation to participants (see Hoyle, Harris, & Judd, 2002, Chapter 12).

Level 1				Level 2			
Participant 1	Participant 2	Participant 3	Participant 4	Participant 5	Participant 6	Participant 7	Participant 8
X_{11}	X_{12}	X_{13}	X_{14}	X_{25}	X_{26}	X_{27}	X_{28}

Figure 3.1a. A between-participants factor defines groups of participants exposed to different levels. X_{ij} is an observation from participant j in condition i .

		First Observation	Second Observation
Group 1	Participant 1	X_{11}	X_{21}
	Participant 2	X_{12}	X_{22}
	Participant 3	X_{13}	X_{23}
Group 2	Participant 4	X_{24}	X_{14}
	Participant 5	X_{25}	X_{15}
	Participant 6	X_{26}	X_{16}

Figure 3.1b. A within-participants factor exposes each participant to multiple levels of the factor. Participants randomly assigned to Group 1 encounter level 1 then level 2; participants in Group 2 receive the opposite order.

Research designs can be divided into three fundamental categories. An *experimental design* involves a manipulation of at least one independent variable, along with random assignment of participants to conditions. Levels of a manipulated factor are often termed treatments. A *quasi-experimental design* involves manipulation without randomization, for example, the assignment of two preexisting groups to different treatments. A *nonexperimental* or *passive observational design* (sometimes confusingly termed correlational) includes neither manipulation nor random assignment. This chapter focuses almost exclusively on experimental design, for two main reasons. The most obvious reason is the heavy predominance of experimental designs in social psychology today. The other is the increasing recognition of the advantages of experimental design, even among researchers who study the effects of social interventions in field settings (Paluck & Green, 2009; see Cialdini & Paluck, Chapter 5 in this volume). Cook and Shadish (1994, p. 557) described “a powerful shift in scientific opinion toward randomized field experiments and away from quasi-experiments or nonexperiments. Indeed, Campbell and Boruch (1975) regret the influence Campbell's earlier work had in justifying quasi-experiments where

randomized experiments might have been possible.”

In social psychology, the units studied are usually individual people, and in this chapter they are called participants. However, the units could equally well be couples, groups, and so on. And the laboratory is usually assumed as the research setting; again, this emphasis simply reflects the predominance of research within social psychology today. Useful discussions of field research (with an emphasis on the practical issues of conducting experimental research outside the lab) can be found in Cialdini and Paluck ([Chapter 5](#) in this volume), Reis and Gosling (2010), Shadish *et al.* (2002), and West, Cham, and Liu ([Chapter 4](#) in this volume).

Independent Variables (Factors)

In an experimental design there are always at least two *factors* or *independent variables*, variables that are considered to influence the dependent variable response. There is a factor corresponding to participants (or whatever other units are randomly assigned) and at least one other factor as well. A factor may be *between participants*, which means that distinct groups of participants are exposed to the different levels of the factor (see [Figure 3.1a](#)). If the participants are randomly assigned to these groups, then the design is an experiment. Randomization means that differences between groups on the dependent variable cannot readily be attributed to alternative causal influences like differences in the composition of the groups and are most reasonably attributed to the treatment itself. Alternatively, a factor may be *within participants*, meaning that each participant is exposed to more than one level of the factor. In this case the order of exposure is what is randomly assigned in an experimental design. As [Figure 3.1b](#) shows, participants are randomly allocated to groups that receive the levels of the factor in different orders. If order is not randomly assigned, alternative causes like learning or maturation could be responsible for systematic differences between measurements taken at different times.

In the following section, we first discuss a series of questions that must be answered regarding each individual factor in a design. Most obviously, how should the levels of the factor be chosen? Then we consider the possible ways that multiple factors included in a single experiment may be interrelated.

Fixed or Random Factor?

Definitions.

The levels of a factor can be considered to be fixed or random. With a *fixed* factor, the levels used in a study are selected for their intrinsic theoretical or practical interest. Any conclusions resulting from the research apply only to those levels. So, for example, in a study of effects of gender stereotypes on participants' evaluations of essays supposedly written by another student, essay writer's gender has two levels, male or female. In a study of the effects of group discussion on decision making, the participants' decisions would be measured before or after a group discussion. With a fixed factor, one wants to use the observations from the sample of participants in each condition to estimate the mean of a hypothetical population of participants exposed to that condition, and then to test hypotheses concerning differences among condition means.

With a *random* factor, in contrast, the research interest goes beyond the particular levels of the factor included in a study. Instead, the hypotheses of interest involve a population of conditions or levels of the factor. The levels used in the experiment are regarded as a sample from the population. The most obvious example of a random factor is the participants factor in a typical experiment. Statistical significance of the results of an experiment indicates that we can safely generalize beyond the specific participants to other “generally similar” participants.

Participants and Other Random Factors.

Consider the way researchers typically treat the participant factor as a basis for conceptualizing other types of random factors. Suppose we wish to investigate the effects of mild versus severe consequences of an accident on people's attributions of responsibility. Scenarios could be written involving various types of accidents (e.g., an auto accident, a kitchen fire), with the severity of the consequences varied. Different participants would read each scenario and make ratings of responsibility for the accident. Results might show that – across all the types of accidents tested – significantly more responsibility is attributed when consequences are more severe. What would that result mean? The answer hinges on the treatment of the “scenario” or “accident type” factor. If the factor is considered to be fixed, then the result can be said to hold for these specific accident scenarios only. The unsatisfying nature of this limited generality can be easily seen from the typical absence of any reference to it in authors' discussions of the implications of their research results. If the factor is considered to be random, however, we can take the result as holding for the population of

accident types similar to those that were tested. This is, roughly, the same type of generalization we expect (and routinely make) across participants.

Neither for participants nor other factors (e.g., types of accidents) is formal random sampling necessary for the treatment of the statistical factor as random. Random sampling, as used in survey research, offers statistical assurance that results from a sample can be generalized back to the parent population, plus or minus the practical problems of actually attaining a perfect response rate and so forth. But in actual research practice we do not randomly sample our participants from among the university's undergraduate students. For one thing, the license to declare that our results generalize statistically to the Midwestern University student body as a whole adds nothing to the interest value of the result and would be purchased at a very high cost. Instead, we rely on independent replications of major results (ideally not just within a single lab but across universities and even nations) as the chief assurance of their generality across participants.

These considerations about participant sampling suggest two lessons for other factors across whose levels we wish to generalize. First, formal random sampling is not a necessary precondition for treating a factor as random. The researcher may write stories about diverse types of accidents rather than sampling from some universe of accident types (even if such a thing existed). In general, we construct and select experimental materials to be (subjectively) particularly good examples of the domain under investigation, rather than typical or representative examples. But such selection is not inconsistent with considering the factor as random. Second, when research is replicated, different levels of the factor should be used (just as different participants would be used) as the most effective way to demonstrate the generalizability of the results across such levels. I will return to this issue shortly. The significant point is that when participants or any other factor is considered to be random, the generalization is to participants or levels of the factor that are similar to those used in the study. Random sampling from a precisely defined population is not necessary to consider a factor as random.

Deciding on the Treatment of a Factor.

When a factor (such as accident type in an attribution study, or the topic of persuasive messages in an attitude change study) could be treated as either random or fixed, how can the researcher decide which to use? As a guideline, consider how one would ideally replicate the study. If one would leave the levels

of the factor the same, the factor is fixed; if one would use other levels that are similar to (not identical with) those used earlier, it is random. Certainly a replication would use new participants, so the participants factor is invariably considered random. Ideally, if researchers recognized that topics and similar factors are arbitrary and conceptually unimportant contexts across which generalizability is desired, they would select new topics in a replication study, which corresponds to treatment of the factor as random.

However, in actual practice, existing persuasive messages, accident scenarios, and similar materials are often reused in replications on the grounds that they are known to “work” (as well as because copying them involves less effort than constructing new materials). This practice is dangerous. For example, early research on the “risky shift” was based on an observation that with certain types of decision problems, group discussion led participants to move toward riskier choices (Bem, Wallach, & Kogan, 1965; see Brown, 1986). Researchers consistently used the same set of problems, which “worked” in the sense of replicating the conventional result. Yet even the original set of problems included a few that tended to produce conservative shifts after discussion. Only when a broader sampling of the domain of possible decision problems was undertaken could it be recognized that the shift to risk was specific to particular problems. The whole phenomenon could then be reconceptualized much more generally (and fruitfully) as group polarization (Myers & Lamm, 1976).

As Abelson (1995) pointed out, a given factor frequently has both fixed and random aspects. His example was a study of the effects of persuasive communications from expert versus nonexpert sources, with messages on several different topics. From one viewpoint, topics could be considered a random factor; the researchers presumably wish their findings to be generalizable across all kinds of controversial topics. From another viewpoint, specific topics could be considered fixed (and in fact particular topics, such as the institution of comprehensive exams for college seniors, have become fixtures of the research literature on persuasion) but the particular messages giving arguments on those topics could be random. That is, a researcher wishing to replicate the initial study could (a) reuse the original materials without change, (b) write new essays on the same topics, or (c) write essays on a completely new set of topics. The original factor is actually topic-plus-essay, which can be seen as a mixture of fixed and random characteristics. Still, the way the factor is treated in the data analysis constrains the generalization that is possible for the research results.

Power with Random Factors.

The greater generalizability that is possible for a factor treated as random obviously must have a cost. Generalization for a fixed factor involves treating the experimental observations as samples that yield evidence concerning the true population means in the given conditions. Generalization for a random factor involves the same inferential leap (sample values to population means for the conditions included in the study) plus another step, treating the factor levels in the study as a sample from a larger population of levels whose overall mean is to be estimated. The latter must involve less power, often drastically less (see Abelson, 1995). The main implication for researchers who wish to use random factors in their designs is to use many levels. Power depends on the number of levels of any random factor, including items (e.g., stories) as well as participants. Typical research practice, strongly influenced by tradition rather than explicit power analysis (Cohen, 1990), involves the use of many participants (perhaps 60–120 in a typical social psychological experiment) but a much smaller number of levels of other factors – often just 2–4. In the case of a fixed factor, this may be fine, depending on the specific hypothesis being tested. But in the case of a random factor, such as stimulus replications, the use of only a few levels of the factor may lead to very low levels of power. Fortunately, it may be easier and less costly to double or quadruple the number of levels of such a factor than to multiply the number of participants by the same factor.

Take the factor “experimenter sex” as an example. This seems on its face like a fixed factor, but the individual experimenters who instantiate the levels of sex are best treated as random. Most researchers realize that if they just have John and Joan serve as experimenters, it is impossible to draw conclusions about effects of experimenter sex, for it is confounded with all of John and Joan's other personal characteristics. But using two or three experimenters of each sex is almost as weak; an analysis that treats experimenter-within-sex as random will have fatally low power. Considerably greater power will result with 10 to 15 experimenters of each sex, even with the same total number of participants divided up among experimenters.

Extremity of Levels

In writing accident scenarios with mild and severe consequences, the exact levels of severity are arbitrary. Severe consequences might include severely injuring one person, or killing three people, or even hitting a gas pipeline and setting off an explosion that kills dozens. In other studies a researcher might

need to manipulate levels of the communicator's physical attractiveness or expertise or the strength of the arguments used in a study of persuasion. How can one decide about the appropriate levels of extremity for manipulations in research?

Matching Naturally Occurring Treatments.

One approach to answering this question, termed “ecological design” (Brunswik, 1955), emphasizes constructing manipulations that match those found in everyday life. Stories involving severe accidents might be sampled from newspaper reports, for example – although a thoughtful researcher might realize that the newspaper reports themselves are a biased sample of the accidents that actually occur. For manipulations of many theoretically specified constructs it is not clear how to apply the principle of ecological design. Should one select communicators whose physical attractiveness falls one standard deviation above and below the mean on some scale of rated attractiveness in a given population?

One advantage of ecological design is that an experiment might yield useful information about the size of a given treatment effect, so that one can judge whether it is practically important or meaningful. This may even enable conclusions about which of several independent variables has stronger effects. A researcher might design a study manipulating two variables X and Y , find a larger effect of X (perhaps no significant effect of Y at all), and declare that X is a more powerful influence than Y on the dependent variable. Under most circumstances this claim is unjustified: Because the extremity of manipulation is ordinarily an arbitrary choice, the researcher might have obtained different results simply by manipulating X less strongly and Y more strongly. However, if the strength of the manipulations was based on naturally occurring variation in X and Y , the conclusion about relative effect sizes may be justified.

Another claimed advantage of using ecologically representative manipulations is an improved ability to generalize experimental results to nonexperimental settings. This idea has a certain intuitive appeal. However, consider the distinction Aronson, Ellsworth, Carlsmith, and Gonzales (1990) made between *mundane realism* (a match on superficial features between an experimental manipulation and some aspect of everyday life) and *experimental realism* (the impact and meaningfulness of a manipulation for the subjects). Ecological design calls for high levels of mundane realism. However, as Brewer and Crano (Chapter 2 in this volume) argue, construct validity depends more on experimental than on mundane realism. And construct validity is the chief

concern with respect to the generalizability of theory-testing research (see also Mook, 1983).

Powerful Manipulations.

The potential advantage of strong (even unrealistically strong) manipulations is statistical power. As long as they do not provoke ethical problems or excite suspicion or ridicule, factors can often be manipulated powerfully in the laboratory. Thus a researcher might try to write extremely strong and extremely weak arguments and choose photos of extraordinarily attractive and unattractive individuals to manipulate the characteristics of the communicator. Strong manipulations will tend to create larger effect sizes (increasing the mean differences among conditions without increasing the error variance that serves as the denominator of the effect size estimate) and more statistical power. It is important to recognize, however, that research manipulations are not invariably more powerful than conditions found in everyday life. Effects of life-threatening illness, the death of a close relative, or watching thousands of murders and violent assaults on television over many years are important to study but cannot be reproduced as manipulations.

Number of Levels

How many levels of a given factor should be used? First, as already noted, if a factor is treated as random, the number of levels effectively becomes the “*N*” that heavily influences statistical power. For example, there may be many stories, scenarios, or videotapes (vehicles for a given manipulation) that serve as replications. In this situation, the more levels, the better.

Another common case is a factor that represents an ordered continuum (more or less severe damage from an accident, strong or weak arguments, stereotype-consistent or stereotype-inconsistent behaviors). The typical research design uses just “high” and “low” levels of such a factor, an approach that maximizes simplicity and statistical power to detect a hypothesized linear effect (McClelland, 1997). However, it can be argued that using more than two levels spaced along the continuum can yield unexpected findings – such as a nonlinear or even nonmonotonic effect of the variable – that may be conceptually important. The cost is reduced efficiency for detecting the linear effect. For example, McClelland (1997) showed that relative to using just the extreme ends of the continuum (with half of the participants in each group), dividing participants into three equally spaced groups gives an efficiency of .67, and five

equally spaced groups just .50. The total number of participants would have to be boosted by 50% or 100%, respectively, to compensate for this loss of efficiency. If the linear effect is of key interest and the possibility of a nonlinear effect is considered only as an afterthought, researchers might reasonably be unwilling to accept this loss of power.

Sometimes researchers make the mistake of dichotomizing continuously measured variables (such as pretests or personality or attitude scores) to fit the overall analysis into a rigid analysis of variance (ANOVA) framework. It is well known that dichotomizing a continuous variable throws away valid variance and reduces power. Still, some researchers argue that if they obtain significant effects of the dichotomized variable despite the acknowledged loss of power, there is no harm done. Maxwell and Delaney (1993) showed, however, that this approach can lead to bias, even making another factor in the design appear to have a significant effect when in reality it does not. There is no good reason to commit this sin in the first place; any regression analysis program can use measured continuous variables as well as classification variables (e.g., manipulated factors) in a single analysis.

Relations among Factors

Multiple factors in a design can be crossed, nested, or confounded. *Crossed* factors include each level of one factor in combination with each level of the other (e.g., strong or weak arguments delivered by an attractive or unattractive communicator, a total of four different conditions). *Nested* factors involve multiple levels of a subsidiary factor that occur uniquely within a given level of a more encompassing factor (e.g., several individual experimenters nested within experimenter sex). Figure 3.1a shows an example, with participants nested within levels. *Confounded* factors are two or more conceptually distinct variables that covary perfectly so their effects cannot be empirically disentangled (e.g., individual experimenters and experimenter sex when there are only one male and one female experimenter).

Reasons for Using Between-Participants versus Within-Participants Factors

One important aspect of a design is whether the factor representing participants (or other units of study, such as groups, couples, etc.) is crossed by or nested within other factors. A between-participants design is one in which participants

are nested within all other factors; that is, a distinct group of participants is exposed to the condition created by each combination of the other design factors. A within-participants design has one or more factors crossed by participants, so that each participant is measured under more than one condition. A *mixed* design has both between-participants and within-participants factors.

The reasons for preferring to manipulate a factor within or between participants generally boil down to power versus the possibility of bias. Power is often greater for a within-participants manipulation, because with many types of dependent variables, differences among participants are by far the largest source of variance. Such differences inflate the error term for testing effects of between-participants factors, but are irrelevant for within-participants comparisons. Sometimes this advantage is stated as “each subject serves as his or her own control” in a within-participants design.

However, within-participants designs allow for several types of bias. *Carryover effects* occur when a participant's response in a given condition depends on experimental conditions that participant experienced previously. This could occur, for example, if previous stimuli serve as an anchor or context for judgments and reactions to the next stimulus. Likewise, previous conditions could create fatigue, prime specific cognitive representations, or influence participants' mood. Fortunately, carryover effects can be detected in the data analysis as differences in responses when conditions are administered in different orders.

In addition, participants in a within-participants design see more than one condition and are thus in a better position to guess at the experimental hypotheses. The resulting *demand characteristics* (Orne, 1962) are an important potential source of bias, as participants start wondering “what are they getting at here” or “what am I supposed to do in this experiment” rather than simply performing the task. For example, a participant who is asked to evaluate the credentials of several “job candidates” may notice that some are men and others are women and leap to the conclusion that the study concerns gender stereotyping and discrimination. It is virtually impossible that this participant's responses to the stimulus persons could remain unaffected by this thought. A participant who evaluates only one candidate is less likely to draw the same conclusion, although the possibility obviously still remains.

Sometimes researchers informally apply an additional consideration, whether the factor of interest or its conceptual relatives occur “between or within participants” outside the laboratory (cf. Brunswik, 1955). For example, an

investigator studying people's reactions to different types of advertisements for consumer products or political candidates might well conclude that in everyday life, people often encounter many ads in quick succession and that the experimental design might as well match that reality.

Reasons for Crossing Factors: I. Testing Theoretically Predicted Interactions (Construct Validity)

Perhaps the most important reason for including crossed factors in an experimental design is to test a predicted interaction. Suppose one finds that people better recall expectation-inconsistent than expectation-consistent information about a person (Hastie & Kumar, 1979). If a theoretical explanation holds that the effect depends on effortful cognitive processing, one might then predict that people who (as a measured individual difference) enjoy thinking hard might show the effect more strongly than other people do. Or one might predict that people under time pressure would not be able to engage in sufficient thought to show the effect, compared with people who have as much time as they need. Such interaction or *moderation* predictions (which can be concisely stated as: time pressure moderates the effect of expectation consistency on recall) have been tested and support the validity of the theoretical explanation. In general, theories that predict more details of a data pattern (such as a specific interaction pattern rather than a simple main effect) gain more in credibility when their predictions are confirmed. This use of interactions in a design improves construct validity. In social psychology, history and tradition seem to give special honor to nonobvious predictions of all kinds, but particularly interaction predictions (Abelson, 1995).

Demonstrating Theoretically Predicted Dissociations.

Theoretical predictions of null effects can be important. For example, if a particular judgmental bias is thought to be automatic and uncontrollable, one might predict that it will be unaffected by participants' motivation to be accurate (manipulated, say, by offering a financial reward for accuracy). Such predictions in general should not be tested standing alone, but in a larger design in which the predictions take the form of an interaction. To illustrate, suppose that one predicts that variable *X* does not influence *Y*. If one simply manipulates *X* and measures *Y*, a finding of no effect is virtually meaningless; too many alternative explanations (including a weak manipulation of *X* or an insufficient *N* resulting in low power) are available. The problems are greatly lessened if one can show

in the same study that X affects Z while not affecting Y . This pattern is termed a single dissociation, conceptually equivalent to an interaction of the X manipulation by measure (Y vs. Z). Still, however, one alternative explanation remains open – that the Y measure is insensitive or resistant to change for theoretically irrelevant reasons. If nothing could affect Y , the theoretical import of the finding that X in particular does not affect Y is minimal. The answer to this interpretive ambiguity is the full *double dissociation* design: manipulate both X and an additional variable W and demonstrate that X affects Z (and not Y) while W affects Y (and not Z). The logic of convergent and discriminant validity, as applied in this form, permits meaningful interpretations of the obtained null effects (see Dunn & Kirsner, 1988).

Interpreting Interactions versus Condition Means.

One interpretive issue involving interactions will be briefly mentioned here: Should interpretation focus on the interaction effect per se, or on the simple effects? For example, in a 2×2 design (say, Participant Sex \times Treatment/Control), should one focus on interpreting the effects of treatment separately within each sex, or on interpreting the interaction effect itself? This issue has aroused some controversy (e.g., Abelson, 1996; Petty, Fabrigar, Wegener, & Priester, 1996; Rosnow & Rosenthal, 1995, 1996). The key is to understand that the interaction effect is precisely a test of the difference between the two simple effects. The direction and significance of each of the simple effects is a separate question. For instance, a significant interaction is consistent with a pattern in which one main effect is near zero and the other positive, or one main effect is negative and the other positive, or even with neither main effect differing significantly from zero. In most research situations, the hypothesis tested by the interaction – that the two simple effects differ – is the hypothesis of interest. More complex hypotheses (such as predicting that one simple effect is positive and the other zero) involve not only the interaction but also the direction and relative magnitude of overall main effects in the design and should probably be avoided.

Power for Interactions.

In an experimental design with equal numbers of participants assigned to the cells, tests of interactions have the same power as tests of main effects. This can easily be seen in a 2×2 design because interactions and main effects can be viewed as different 1-*df* contrasts on the four condition means. However, when

one or more factors are measured rather than manipulated, statistical power for testing interactions can be lower, sometimes by an order of magnitude, than power for testing main effects in the same study. First, if the design includes few cases that have extreme values on both of the crossed factors, power for testing the interaction can be quite low (McClelland & Judd, 1993). Of course, this may be the case for variables that are personality scores or other more or less normally distributed measurements, but not if the two crossed factors are experimentally manipulated. Second, low power for tests of interactions can result from measurement unreliability. If the two factors that enter into the interaction are imperfectly measured, it is well known that tests of their effects lose power – and tests of their interaction lose even more (Busemeyer & Jones, 1983; Jaccard & Wan, 1995). Again, experimentally manipulated factors (which generally have no measurement error) do not suffer from this problem. But in a field study or in an experiment that includes one or more measured independent variables, low power for interactions is a serious threat to a researcher who is theoretically interested in interaction effects.

Data Transformations Complicate Interpretations.

Another issue concerning interactions is that with ordinal interactions (those that do not involve crossover patterns), mathematical transformations of the dependent variable can create or remove interactions. For example, if the four means in a 2×2 design are 1, 4, 4, 9, there is an interaction (the last cell is higher than it would be with only two main effects). But applying a square-root transformation results in means of 1, 2, 2, 3, displaying no interaction. For this reason, unless the applicability of transformations can be ruled out on a theoretical basis, claims about the presence or absence of an interaction are weak. That is, unless it can be argued that a log, square-root, square, or some other transformation cannot meaningfully be applied to the response scale, it is difficult to conclude whether an ordinal interaction is present or absent in the data. As an example, several studies have drawn theoretical conclusions based on findings suggesting that long-term (chronic accessibility) and temporary (priming) sources of accessibility of mental constructs do or do not interact (e.g., Bargh, Bond, Lombardi & Tota, 1986). But the data patterns never show crossovers, and so conclusions as to the additivity or interactivity of these sources of accessibility are weak at best. No argument against the application of transformations is likely to be convincing. The level of activation of an unobservable mental representation (the construct of true theoretical interest) has an effect through two steps, neither obviously linear: Activation affects the

probability or intensity of use of the mental representation to encode a given stimulus, which in turn affects an overt rating scale response when the participant is asked to make judgments about the stimulus. In cases like this, in which the assumption of linearity of the construct-response scale relationship is indefensible, an argument can be made for applying whatever standard data transformation makes the results of a body of studies simplest – for example, main effects without interactions (Abelson, 1995).

Reasons for Crossing Factors: II. Reducing Error Variance (Statistical Conclusion Validity)

Besides increasing construct validity by testing specific theoretical predictions, researchers may have other reasons for including crossed factors in a design. A second reason is related to statistical conclusion validity: Added factors can reduce error variance and therefore increase the statistical power of a design to detect an effect of interest. In general, variables affecting a particular dependent measure can be held constant, allowed to vary unsystematically, or varied systematically. For instance, if the researcher is interested in the effect of strong versus weak persuasive arguments on attitudes concerning a controversial social issue, he or she may attempt to hold constant participants' motivation to read the persuasive message carefully (by instilling a high level of motivation in all participants). Differences in the amount of participants' knowledge about the issue may be allowed to vary unsystematically. And participant factors such as gender, or situational factors such as the time of day, may be systematically varied and recorded so that their potential effects on the dependent variable can be assessed.

Covariates and Power.

Relative to allowing an influential variable to vary freely, either holding it constant or statistically controlling for its effect will decrease the amount of residual error variance in the dependent variable. In turn this increases the statistical power of the study to detect effects of other independent variables of interest. Factors that are included in an experiment specifically for this purpose, because they represent theoretically irrelevant sources of variance that are known to have strong effects, are termed covariates (if they are continuous measured variables) or blocking factors (if they are categorical). For example, a pretest attitude score could be used as a covariate. If the measure is continuous and has a linear relation to the dependent variable, power is maximized by

treating it as a covariate; if it is categorical or has a nonlinear relation to the dependent variable, using it to create blocks is preferable (Feldt, 1958).

Analyses with Covariates.

In the results of such a study, a main effect of a covariate or blocking factor (e.g., a pretest score) may be of little theoretical interest. Interactions between the factor of conceptual interest and the covariate or blocking factor, however, may be theoretically important. For example, a given message may be found to have a larger persuasive effect on participants with one initial attitude position rather than another. Such interactions have implications for the generality of the treatment effect and are discussed more fully later in this chapter. Treatment-by-covariate interactions are not considered in the traditional analysis of covariance framework, but can be tested without difficulty in a general linear model framework (Cohen, 1968).

Covariates may be measured once per participant or once per observation in a within-participant design. To illustrate, imagine a study in which participants make ratings of their overall impression of a social group as they encounter information about more and more individual group members. In this design, “number of group members encountered” is a within-participants independent variable. If the researcher can obtain participants’ scores on a personality measure, such as authoritarianism, which is expected to influence ratings of out-groups, this variable might be used as a covariate (with a single score per participant). If it is believed that participants’ mood might also affect the positivity of their ratings of the group, an assessment of momentary mood might be taken at the same time each rating is made, and used as a within-participants covariate. Either of these types of covariate might not only have a main effect on the dependent variable, but might interact with the experimental factor. For instance, effects of increased exposure to the out-group might lead to more positive ratings among participants in positive moods. Judd, Kenny, and McClelland (2001) discussed analysis of such designs.

An important conceptual point regarding covariates is the time of their measurement and their conceptual status as control variables or mediators. Ordinarily, covariates (or blocking factors) used purely to control error variance are measured before the treatment is applied. If treatments are randomly assigned, this means that the covariate and the treatment factors are expected to be independent. However, what if the covariate is measured after the treatment? In this case the treatment might affect the covariate as well as the dependent

variable, and controlling for the covariate would mean that the treatment effect would not be properly estimated. However, this exact design is used when one considers the covariate as a potential mediator of the treatment effect on the dependent variable – a situation that is discussed later in this chapter.

Because this chapter focuses on experimental design, it is assumed here that the main factor of interest (such as strong versus weak arguments in the example) is randomly assigned and manipulated. In this context, the use of covariates (measured before random assignment) serves to increase power but is not necessary to correct for initial differences between the groups of participants exposed to the different treatments. Indeed, because of random assignment, such differences are not expected to exist. In a nonexperimental design, often termed the nonequivalent control group design, levels of a treatment are administered to intact or self-selected groups that differ in unknown ways. In this case, covariates serve (more or less unsatisfactorily) to adjust for some of the differences between the treatment groups. This design raises many difficult issues; see discussions by Cook and Campbell (1979) and West *et al.* (Chapter 4 in this volume).

Reasons for Crossing Factors: III. Establishing Generality of an Effect (External Validity)

A third reason for including crossed factors in a design is related to external validity: the desire to establish the generality of a given effect across multiple levels of another factor. Once again, recall that the three ways of handling a potential influence on a dependent measure are to hold it constant, allow it to vary unsystematically, or vary it systematically (manipulate or measure it). We noted earlier that holding constant and varying systematically were preferable from the viewpoint of statistical conclusion validity. But varying unsystematically and varying systematically are preferable from the viewpoint of external validity.

Varying a Factor Unsystematically.

Let us consider three ways of allowing variation in a contextual factor, one that might influence a dependent variable but whose effects are not of primary theoretical interest. First, allowing the factor to vary unsystematically (as compared with holding it constant) offers the advantage of establishing at least a minimal level of generalizability of the experimental effect of interest. For

example, if an experiment calls for a confederate to interact with participants and administer a manipulation, a researcher might use five different confederates (not just one), with confederates assigned haphazardly to run different experimental sessions based on scheduling convenience, *etc.* Such a study can show at least that the result holds averaged across the unsystematic influences of the confederates' personal characteristics (see Cook & Campbell, 1979). However, one cannot say for certain that the effect holds for each level of the unsystematically varied factor.

Varying a Factor Systematically.

Second, systematically varying the contextual factor (e.g., using “confederate” as a design factor crossed with condition, including randomly assigning participants to confederates) offers potential advantages. To the extent that contexts have lawful effects on the dependent variable, this approach, compared with unsystematic variation, will reduce error variance and increase power (see above). And the analysis can potentially establish that the effect of interest holds at each level of the contextual factor (not just averaged across levels, as in the previous approach). The drawback is that the contextual factor probably should be considered as random, as discussed earlier, with important implications for the power of the research.

The ideal pattern of results is an effect of the theoretically important factor that occurs in the absence of any interactions of that factor with the contextual variables (such as confederate). Abelson (1995) discussed issues involved in interpreting such interactions, particularly the differences between qualitative and quantitative interactions. He introduced three useful terms: a *tick* is a specific, articulated finding, such as a reliable difference between two means; a *but* is a qualification on a tick; and a *blob* is an undifferentiated and therefore meaningless “finding” (e.g., a significant multi-*df* *F*-test). Examples may illustrate the use of these terms. “Compared to the control condition, Treatment X improved performance on Task Y” or “...on all tasks” is a tick. “Compared to the control condition, Treatment X improved performance on Task Y but not on Tasks Z or W” – a quantitative (noncrossover) interaction – is a tick and a but. “Treatment significantly interacted with Task” is a blob. “Treatment X improved performance on Task Y but decreased performance on Task Z,” a crossover interaction, is two ticks – really two findings rather than a single finding with a qualification. And “Treatment X improved performance to the degree that subjects expected success on the task” could be an insightful and parsimonious

redescription of the two-tick result as a single tick (Zajonc, 1965 is a classic example). Abelson's introduction of these terms leads to a useful metric for assessing the merit of alternative descriptions of results: Never use blobs, minimize ticks (in the interests of parsimony), and minimize buts (in the interests of generality).

Varying a Factor Across Studies.

A third approach is to hold a given variable constant in a particular study and rely on variation of contexts across studies to establish generalizability of an effect of interest. This approach may be taken either within a given investigator's lab (e.g., use one confederate in one study and a different confederate in a similar study conducted the next semester) or across labs. Replication across labs is obviously the most feasible approach to varying several potentially important types of contexts that might influence a research result, such as the participant population, type of equipment used, specific implementation of treatments, and so on. Replications across different studies, whether in one or many labs or other settings, can be summarized with the technique of metaanalysis (Johnson & Eagly, Chapter 26 in this volume).

Researchers interested in the analysis of social interventions in natural settings (Cook & Shadish, 1994) and those concerned with the generality of theoretical conclusions from laboratory-based research (e.g., Abelson, 1995; Hedges & Olkin, 1985) have recently converged on a recommendation to take this third approach: replicate in separate studies and meta-analyze to establish generality. Abelson (1995) saw a narrow focus on generalization from a single study as misplaced, given that, “In practice, real research investigations usually involve multiple studies conducted as part of an ongoing conceptual elaboration of a particular topic. This permits a richer variety of possible assertions about outcomes” (p. 38). Similarly, Cook and Shadish (1994) noted:

Interest has shifted away from exploring interactions within individual experiments and toward reviewing the results of many related experiments that are heterogeneous in the times, populations, settings, and treatment and outcome variants examined. Such reviews promise to identify more of the causal contingencies implied by [the philosopher] Mackie's fallibilist, probabilistic theory of causation; and they speak to the more general concern with causal generalization that emerged in the 1980s and 1990s.

(p. 548)

Additional Considerations.

There are several other considerations involving replication. First, the emphasis here is on what has sometimes been called conceptual replication rather than exact replication. In a *conceptual replication*, the researcher looks for new ways to manipulate or measure the conceptual variables of interest, rather than striving to reproduce exactly all the procedures of the original study. Exact replications are rarely carried out in social psychology; the most important exceptions are when a study is being criticized. For example, conducting an exact replication study seems to be necessary when a new researcher believes the original results are simply a product of Type I error, or when he or she wishes to demonstrate that some confounding factor produced the results of the original study and that removing the confound will change the results. Outside of these limited circumstances, replications ordinarily involve some variation rather than precise following of a recipe, and they often involve the addition of extra conditions or measures in an effort to explore the limits of an effect as well as incorporating the conditions of the original study.

Second, researchers frequently overestimate the likelihood of a replication being successful. After all, the original study demonstrated the effect, so shouldn't a similar study be able to also? However, Greenwald, Gonzalez, Harris, and Guthrie (1996) showed that if an original study produced a significant effect at the .05 level, the chance of obtaining significance in an exact, independent replication using the same N is quite low. Only if the first study produced $p < .005$ does the replication have a power level of .80 to detect the effect at the .05 level! This is an instance of researchers' general overoptimism about power levels and their consequent tendency to run low-powered studies, which has frequently been noted and decried (Cohen, [1962](#), [1990](#)).

Nonindependence of Observations in Within-Participants Designs

In any design in which multiple measurements are taken per participant (or other unit, such as couple or group), the observations must be assumed to be nonindependent. For example, evaluations of multiple stimulus persons given by a single participant will probably be positively correlated, as a result of the participant's expectations about people in general, tendencies to use particular ranges of the response scale, and so forth. In other situations, such as in a study

measuring the amount of talking time by leaders versus other members in a problem-solving group, the variables are likely to be negatively correlated: the more talking by one person, the less by others. Hidden sources of nonindependence may be present and contaminate observations even when a design is intended to focus on individual participants. For example, if several participants at a time are seated in a room to fill out individual questionnaires, they may influence each other (perhaps by muttering comments about the questions or their answers), or situational factors (such as an uncomfortably warm room or an audible disturbance in the hallway outside) may influence all participants in a group.

If nonindependence between observations is ignored in the data analysis, bias will result. Judd and McClelland (1989) described the direction of bias in different situations. When participants (or groups, or other units that produce multiple observations) are nested within conditions, a positive correlation produces a “liberal” bias (F statistics too large) and a negative correlation produces a “conservative” bias (F statistics too small). When participants are crossed with conditions, the opposite pattern emerges. Thus, nonindependence of observations can result in an increased likelihood of either Type I or Type II errors depending on the specific circumstances. The bias can be large in realistic circumstances (Kenny & Judd, 1986).

As a result, the data analysis must take account of nonindependence. Two approaches are widely used: repeated-measures ANOVA and hierarchical linear models (HLM, also termed multilevel models). The latter is more general; for example, it can handle situations in which different participants have differing numbers of observations (through measurement lapses or by design). The newer HLM approach offers other advantages as well. The issue goes beyond the scope of this chapter, but see Gelman and Hill (2007, Chapter 11).

Counterbalancing and Latin Square Designs

Counterbalancing.

Suppose that one wishes, for theoretical reasons, to investigate the effect of a prior expectation about a person on recall of the person's behavior. One could tell participants that “John” is an honest fellow and then expose them to a list of some of his honest and dishonest behaviors (randomly ordered). Hastie and Kumar (1979) performed a study like this and found superior recall for the expectation-inconsistent behaviors. However, there is a problem: Honest and

dishonest behaviors necessarily differ in content, so one set might be more unusual, more distinctive, more concrete, more imageable, or different in some other way that affects memorability. Such uncontrolled factors (and not inconsistency with expectations) might have led to the superior memory.

There are two possible approaches to this type of confounding. One could attempt to have judges prerate all the stimuli (here, the behaviors) for concreteness, distinctiveness, imageability, and so on, and then select sets of honest and dishonest behaviors that are exactly equated on all these factors. But this approach, beside being incredibly cumbersome, is ultimately doomed to failure: However long the list of factors that can be rated and equated, it can never be guaranteed to include all that might affect the items' memorability. Hastie and Kumar (1979) chose an alternative approach that is both simpler and more effective. They employed a *counterbalanced* design, in which the same behaviors served as expectation-consistent stimuli for some participants and expectation-inconsistent ones for others. That is, different groups of participants (randomly assigned) initially learned that John was honest or that John was dishonest. Then all participants saw the same list of behaviors. This approach uses the logic of design to ensure that all factors affecting the memorability of a behavior – even those that the researchers are unaware of – are equated between consistent and inconsistent behaviors, across the entire experiment.

As another example, a researcher may need to construct stimulus materials that are specifically tuned for each participant, using words or constructs that a particular participant employs frequently (termed “chronically accessible”) and also words or constructs that are employed less frequently (e.g., Higgins, King, & Mavin, 1982). A solution is pairing participants and using A's chronically accessible words as the less accessible words for B, and vice versa. Across the entire experiment this design guarantees that the distribution of characteristics of accessible and less accessible words is the same.

Pairing participants in these ways creates nonindependence within each pair of participants, requiring that pairs be treated as units in the analysis. For example, the characteristics of A's accessible words influence both A's and B's responses in the study. In addition, one important consideration is whether there is some clear basis for assigning the members of each pair to the “A” and “B” roles in the analysis. A study of dyadic interaction, for example, may pair a male with a female or a parent with a child in each dyad, or may bring together pairs of male participants. In the latter case, with nothing in particular to differentiate the participants, results can be affected by the way they are assigned to roles in the

analysis; see Griffin and Gonzalez (1995) or Kenny, Kashy, and Cook (2006).

Latin Square Designs.

Counterbalancing frequently makes use of a type of design termed a Latin square. Figure 3.1b shows the simplest instance of this design, with two groups of participants who receive two levels of the treatment, in a counterbalanced order. Group 1 receives level 1 then level 2, and Group 2 receives level 2 then level 1. Analysis of such a design (see any ANOVA textbook or Judd & McClelland, 1989) can separate order effects from treatment effects. Note one important qualification, however: Treatment is completely confounded with the Group \times Order interaction. Thus, treatment effects can be interpreted only if it is assumed that this interaction is absent. In an experimental design, where participants are randomly assigned to groups, this assumption poses no problem: Randomly constructed groups of participants should not differ in the effect of order on their responses. Nevertheless, in principle, this interaction is confounded with treatment; similarly, Group \times Treatment is confounded with order. The Treatment \times Order interaction (confounded with group) is the one to examine for the presence of any carryover effects (e.g., effects of a treatment that differ depending on whether it was the first, second, or any subsequent treatment encountered by a participant).

The same principle can be further generalized to the situation where factors other than order are associated with different treatments. In many studies, a treatment is embedded in some type of stimulus that serves as a vehicle (e.g., a photo manipulation of physical attractiveness embedded in a folder of information about a target person, a manipulation of source expertise embedded in a persuasive message about some topic). For reasons of power, it is often desirable to use within-participants designs with such materials. However, a given participant can see each stimulus in only one condition (it would be nonsensical to see different photos paired with the same person information, or the same message attributed to both an expert and nonexpert source). Again, a Latin square design is the solution.

Kenny and Smith (1980) presented methods for constructing and analyzing appropriately counterbalanced designs, where the rows are groups of participants and columns are groups of stimuli, and treatments are assigned to particular cells of the design. Kenny and Smith (1980) emphasized that the stimulus factor (e.g., specific stimuli) is properly considered as random. As Santa, Miller, and Shaw (1979) have demonstrated, simply summing over stimuli to form a dependent

measure for each condition, or mistakenly treating the stimuli factor as fixed, can give rise to upward-biased F -tests. This is because variation attributable to items or stimuli would be confounded with variation attributable to conditions. This error frequently occurs in the literature.

Nested Factors

In addition to being crossed or confounded, factors may be nested. We have already discussed the issues involved in the most common cases of nesting, where the inner (nested) factor is random (e.g., participants within experimental conditions, individual experimenters within experimenter sex). Rarely, a nonrandom nested factor is used, such as a manipulation of specific occupations within occupational status ([low status] janitor, secretary; [high status] manager, physician) in a study of occupational stereotypes. In a case like this the researcher may be interested in the specific effects of each level of the nested variable (occupation in the example) and also in comparing the larger categories (low vs. high status). However, this situation need not be conceptualized as two nested factors and in fact usually is not; occupation can be treated as the only design factor and planned comparisons can be performed to examine effects of status.

Dependent Variables

Issues involving the dependent measures in a study are also part of experimental design, and these choices can influence statistical conclusion validity (power), internal validity, construct validity, and external validity.

To Pretest or Not to Pretest?

Pretesting participants with a measure identical or related to the dependent variable can increase power. As discussed earlier, the pretest can be used as a covariate or to create a blocking factor, with blocks composed of participants with similar pretest scores. The power depends on the covariate's within-groups correlation with the dependent variable. If that correlation is r , the experiment's error variance decreases by approximately a factor of $1 - r^2$ in either a randomized-blocks or covariance design (Kirk, 1968). If a covariate with a correlation of .70 is available, its use would cut error variance roughly in half, then, doubling all F statistics.

Yet pretesting is not common in social psychological experiments. Why not? Some obvious reasons include the popularity of types of studies in which pretests hardly make sense, such as social cognition studies in which the dependent variables are various measures of participants' judgments concerning specific stimulus materials. Participants cannot be asked to judge the same materials before the study begins. Likewise, in within-participants designs where variation between participants does not enter into the error term, pretesting is of little use. But even in research on effects of persuasive messages, in which a pretest measure of the target attitude is quite feasible, pretests are often not used. The main reason is a concern about biases caused by exposure to the pretest (Cook & Campbell, 1979). Participants who have completed a pretest measure of some attitude or other construct may respond differently to a relevant manipulation than unpretested participants, weakening the generalizability of research conclusions. Here is yet another trade-off in research design, the possibility of pretest sensitization versus power.

Another approach to avoiding pretest sensitization is to administer a pretest in a separate session, completely unconnected with the main experiment in the participants' minds. For example, in a mass pretesting session at the beginning of an academic term, members of the participant pool may fill out numerous measures. Among these may be pretests for experiments in which they may participate later in the term. This type of procedure should greatly reduce or eliminate the possibility that participants will respond differently in the main experiment than they would have without the relevant pretest.

The Solomon Four-Group Design, which is essentially a Pretest/No Pretest \times Treatment/Control 2×2 , can be used to examine whether pretest sensitization effects exist in any given instance (Campbell & Stanley, 1963). But this design is rarely employed in practice, for if researchers suspect the possibility of pretest sensitization, they usually prefer to avoid it rather than to study it.

Selection of Items for Dependent Measure

One important issue regarding the selection of items or stimuli of any sort in the construction of a dependent measure has strong parallels with an issue discussed earlier, the treatment of "contextual" factors that are not of primary theoretical interest (such as different types of accident scenarios that serve as vehicles for a manipulation of accident severity). In fact, "items" could be considered a within-participant factor. But the issue is reviewed here because conceptually it has to do with the nature of a dependent measure. As in the case discussed earlier, there

are several possible choices a researcher might make. In the following, an “item” could be a question on a pencil-and-paper questionnaire, a stimulus that participants make judgments about, and so on.

Use One Item.

A researcher may use a single item, presumably selected on the basis of pilot testing as “the best.” This is similar to holding constant a contextual factor as described earlier, such as using only a single experimental confederate. This approach has the advantage of minimizing the participant's time and effort, but it has very important disadvantages as well. It may lead to low reliability and little variance, for the resulting measure has only a single item. More important, holding any factor constant leaves the extent of the generality of the findings uncertain.

Average across Multiple Items.

Most frequently, researchers employ several items and average or sum them to form the dependent variable for analysis. This is the most common approach when multiple items have the same scale (e.g., a 1–7 Likert response scale). This approach corresponds to unsystematically varying a contextual factor, such as using several confederates but not treating confederate as a factor in the data analysis. This approach permits a modicum of confidence about the generality of a result, which can be demonstrated to hold across a set of items (though not necessarily for each individual item). This approach also gains power by the use of multiple items. Bush, Hess, and Wolford (1993) described alternative ways to combine multiple measurements within participants, some that may have considerably higher power than simply taking a sum or mean of the raw observations. Even if the sum or mean is preferred for reasons of conceptual simplicity, researchers may wish to standardize each item separately (e.g., to zero mean and unit variance) before combining them, to avoid overweighting items with larger variances.

Treat Items as a Factor.

A researcher may use several items and treat “item” as a within-participants factor in the analysis. Again, each item should have the same scale (either in raw form or after standardization). This approach corresponds to systematically varying a contextual factor, such as using several confederates and analyzing for

their effects. Compared with the use of a single item, this approach gains power by the use of multiple items. It also permits good confidence about the generality of a result if it can be demonstrated to hold across items.

There is a catch, however. “Items” should probably be treated as a random rather than fixed factor (the same argument was made earlier about confederates or similar contextual factors). This means that a large number of items need to be used before the appropriate statistical tests will have adequate power to show that the effect of interest generalizes to the entire population of items similar to those used in the study. (As noted earlier, a random factor need not involve explicit random sampling from a population – for instance, the participants factor is invariably considered random, even though social psychologists rarely perform formal random sampling of participants.) Using a large number of items is essential if studies are to yield conclusions that can be expected to generalize across items (stories, photos, behavior sentences, etc.) rather than holding only for those used in a given study (Abelson, 1995; Kenny & Smith, 1980).

And there is another catch – or what seems to be one. Treating items as an explicit factor in the design (whether fixed or random) allows for the detection of Treatment \times Items interactions. These interactions often raise questions that researchers would rather not face, for they point to inadequacies in our theories or our measures. Yet such effects tell us about the limitations of our theories and can even spur fundamental theoretical advances. Recall that the “risky shift” phenomenon, when it was found to reverse with new types of decision problems, was felicitously reconceptualized as “group polarization” (Myers & Lamm, 1976). Positive effects of an audience on performance of some tasks and negative effects on other tasks gave rise to Zajonc's (1965) model of drive strength and social facilitation. Findings that the strength of persuasive arguments did not affect attitude change equally for all topics played a role in the development of the elaboration likelihood model of persuasion (Petty & Cacioppo, 1979). In all these cases the trick, as Abelson (1995) explained, is to take a Treatment \times Contexts or Treatment \times Items interaction and come up with a creative and insightful theoretical account of the nature of the differences across contexts or items that causes the treatment to behave differently. Such an account, if it passes tests in further studies specifically designed for the purpose, necessarily integrates more data than the earlier theories that predicted general main effects of treatment across all items or contexts.

Analyze Each Item Separately.

Occasionally a researcher may choose to analyze multiple dependent variables separately. Conceptually, two different situations must be distinguished (see Abelson, 1995). If the multiple dependent variables are regarded as alternative measures of the same construct yet produce different results, it is not clear what to make of that. Perhaps a minor deviation in results for one measure is simply owing to measurement error or other chance factors. Certainly one would not want to place any importance on the fact that a given effect reached $p < .05$ for one measure but only $p = .06$ for another. In this situation one should probably avoid analyzing the different items separately but rather analyze with items as a factor and consider a deviation to be meaningful (a “but” in Abelson's terms, a qualification on an overall pattern of results) only if the items factor significantly interacts with one of the experimental factors. If several such interactions emerge, clearly the original assumption that the items all reflect a common underlying construct needs to be rethought.

If different dependent measures are considered to reflect distinct constructs, then analyzing them independently (rather than combining them into a single index for analysis) is a reasonable approach. Parallel results might be described with a phrase like “not only”; for example, “The effect of treatment was found not only for the attitude measure but also for behavioral intentions.” In Abelson's terms, this is two ticks (two discrete findings) rather than a tick and a but (a limitation or qualification on a finding).

Use a Structural Equation Model.

When measures of multiple conceptual variables are included in a study and the researcher has hypotheses concerning their causal and mediational interrelationships, the data can be considered in a multiple-group structural equation framework. For example, a design might have a 2×2 with a manipulated factor (say, treatment vs. control) and a measured independent variable (such as participant sex) and dependent measures of attitude, behavioral intention, and behavior. The researcher might hypothesize that the independent variables will affect the dependent variables in particular ways, and also that the dependent variables have specified causal paths among themselves. Further discussion of this approach to conceptualizing and analyzing an experimental design is to follow.

Use MANOVAs.

For completeness, let us consider one more potential approach to the analysis of multiple dependent measures: MANOVAs. At least if the overall multivariate analysis is not followed up with specific univariate tests, the MANOVA approach in this situation is unsatisfactory (Abelson, 1995). The reason is that a significant multivariate F is a blob that tells nothing about the actual dependent variable that was analyzed; it only indicates that some empirically derived linear combination of the items significantly differentiated the groups. Outside of a purely exploratory context, this finding tells the researcher nothing with any theoretical content.

An exception to this generalization is that, according to Cole, Maxwell, Arvey, and Salas (1993), MANOVA is appropriate for the analysis of causal indicators, a set of observed variables that are regarded as causes (rather than effects) of the conceptual dependent variable. For example, variables reflecting the presence or absence of several diseases or disabling conditions might be combined into a measure of overall health status. Causal indicators differ from the much more common situation in which the observed indicators, like responses to several questionnaire items, are regarded as caused by the underlying construct of interest. (1) Causal indicators actually cause the dependent variable (e.g., a worsening of arthritis causes reduced health status, although a change in one attitude item would not cause a change in an overall attitude); and (2) there is no special reason to expect cause indicators to be correlated (e.g., someone with worse arthritis will not necessarily have worse asthma as well, although multiple indicators of an attitude should be correlated). The use of causal indicators (and therefore this application of MANOVA) appears to be rare in social psychology.

Additional Considerations Regarding Design

Power

Low Power of Typical Designs.

The design-related advice that receives probably the most lip service and the least frequent implementation is advice to consider the power of a study. Cohen (1962) found that published studies in social psychology had a median power to detect a “medium-sized” effect of only .46. This means that even researchers who are clever enough to be testing a hypothesis that actually was true and had a

medium-sized effect have less than an even chance of finding a significant result. Despite much discussion of the issue in the intervening years, more recent surveys of published studies found virtually identical results (Cohen, 1990). And, of course, the population of all studies must have much lower average power than the subset of studies that are published. Because Cohen (1988) provided researchers with easily accessible ways to evaluate the power of a planned or already conducted study, researchers have little excuse for wasting effort by conducting low-powered studies.

Consider the plight of a researcher who develops an interesting hypothesis, tries three times to test it in low-powered studies obtaining nonsignificant results, and gives up in discouragement, when a single study with the same total N (or a metaanalysis of the three studies, or an analysis of the combined data) might have found a significant effect. Principled advice to this researcher would have been to conduct a power analysis before running the first study (using related results in the literature as guidelines to the expected effect size) to determine how large a study to run. At least, after the first study yielded nonsignificant results, a power analysis would be in order, using the actual effect size estimate from that study.

Power Is Not the Only Consideration.

However, in the real world of research, other considerations may temper this advice. For a novel and untested hypothesis, the researcher might not wish to devote the resources to a single adequately powered study (or might not have those resources in the first place), so it might be rational to place a relatively small bet and hope for a favorable outcome despite long odds. (We hope that tenure does not hang on the outcome, or if it does, that the researcher is placing many independent bets.) Also, a researcher might learn things from a small initial study – despite the nonsignificance of the hoped-for main result – that would allow the refinement of the procedure, measures, or manipulations for succeeding studies. Some might argue that a researcher who incrementally modifies the details of a paradigm over several unsuccessful studies and finally “gets it to work” has learned a lot about the parameters of his or her effect.

Against this plausible argument stand two considerations. First, assessment of the value of this knowledge must take into account the fact that it will almost always remain as local “lab lore” rather than being shared widely with the research community through publication. Second and more important, running several studies each with a power of .30 or so to detect a real effect means that

one will eventually pop up as significant by chance alone, even if the procedural variations that are being tested make no real difference. The pattern of results would then convey no actual information about the supposed benefits of the variations, so the “lab lore” would be purely illusory. (And it should be obvious that if the researcher publishes the study that “worked” without even mentioning the several variants that were tried and did not work, the odds that the published result is a Type I error are greatly increased.) The point of this discussion is not that power analysis is unimportant, but that decisions about how to spend research resources (particularly participant hours) are among the most difficult faced by any researcher, and power is only one relevant consideration.

Ways to Increase Power.

Besides the number of participants, other design features that can contribute to statistical power include the choice of within-participant rather than between-participants designs and the use of covariates or blocking factors, reliable measures of dependent variables, powerful and consistent implementations of treatments, and focused one-*df* tests of key hypotheses (rather than multi-*df* omnibus *F* -tests). All of these issues have been mentioned earlier in this chapter.

One additional consideration is the use of optimal design, or the allocation of a fixed number of participants across conditions in a way that maximizes power. McClelland (1997) gives a concise treatment of the issues. Here is an example based on McClelland's discussion. A researcher may wish to test three conditions consisting of a control (C) and two experimental treatments (E1 and E2). For example, in an attitude change study, an unadorned version of a persuasive message may serve as the control, and the message plus two different heuristic cues (such as an attractive source and an expert source) may be used in the experimental conditions. The researcher plans to test two single-*df* hypotheses: (1) comparing E1 and E2 together against C to see if any heuristic cue aids persuasion, and (2) comparing E1 against E2 (ignoring C) to see which heuristic cue performs better. McClelland showed that no allocation of participants across the three conditions is optimal for both of these hypotheses simultaneously. If the first contrast is of crucial importance, then half the participants should be allocated to the C condition and one-quarter to each of the others for maximal power (using random assignment but with unequal probabilities). However, then the efficiency for the second contrast is only .5. On the other hand, if the second contrast is of central importance, all the participants should be divided between

E1 and E2 for maximal power, and the first contrast cannot even be tested. An equal- N division (one-third to each condition) gives higher relative efficiency to the first contrast than to the second. This example (testing two conditions against a common control and against each other) is a common one but the principle is even more general. The efficiency of a design with a fixed number of participants is affected by the way the participants are allocated to conditions, and the allocation that gives maximal power depends on the specific hypotheses to be tested.

Unequal N s

Cell sizes in a design may be unequal when the design reflects a sample from an underlying population in which different categories have unequal sizes (e.g., male and female liberal arts and engineering majors). Or they may be unequal even when an investigator was attempting to produce equal cell sizes as a result of procedural mistakes, participant no-shows, and the fuzziness of randomization. Neither of these situations is a major problem. Both can be analyzed with readily available programs that use the general linear model approach, although one must be wary of the different assumptions used by different programs. Also, modestly unequal cell sizes do not hurt power very much. For example, McClelland and Judd (1993) demonstrated that in a 2×2 design, even when half of the observations fall into one cell with the remainder spread out equally across the other three, efficiency for detecting an interaction effect is 80% of what it would be with exactly equal N s.

Confounds and Artifacts

Some types of confounds cannot be ruled out by the randomization used in an experimental design. These constitute threats to construct validity rather than internal validity (Cook & Campbell, 1979). There is no question that the experimental treatment as manipulated causes the observed effects – the question is whether it operates through the theoretically postulated mechanism or in some other, conceptually less interesting way. Among the potential confounds in an experiment are the following.

Demand Characteristics and Experimenter Bias.

Participants' reactions to their perceptions (correct or incorrect) of the purpose of the research may influence their responses. Standard precautions against

demands include the use of a coherent and believable cover story, which can at least ensure that all participants hold the same views of the purpose of the experiment rather than varying in their perceptions (perhaps in a way that is correlated with conditions). Some researchers apparently believe that if participants cannot correctly guess the research hypotheses, demands could not influence their behavior. This belief is mistaken, for even incorrect guesses may well influence participants' behavior as they attempt to be good participants by confirming what they consider to be the hypotheses (or bad ones by disconfirming them).

Another precaution is to keep experimenters, confederates, and others who have contact with participants unaware of each participant's condition as much as possible. For example, an experimenter might administer a treatment to participants and then immediately leave, with the dependent measures collected by a different experimenter (who is kept blind to condition) or by computer. Such precautions mean that unconscious biases cannot cause the experimenters to treat participants in subtly different ways in different conditions, which might artifactually lead to confirmation of the experimental hypotheses. What is sometimes regarded as an alternative – keeping experimenters unaware of the experimental hypotheses – offers no real protection against biases that vary by condition (for the experimenters will come up with their own, possibly misguided, ideas about the research hypotheses).

Differential Treatment-Related Attrition.

If participants drop out from the experiment whether by physically leaving or by failing to follow instructions, giving nonsensical answers, and so on, it may falsely produce differences between conditions. The only real solution is to have no attrition (see McClelland, Chapter 23 in this volume). It is sometimes assumed that if attrition rates are relatively equal across conditions, there is no problem, but this is incorrect; in one condition, 5% of the highest self-monitors (or most motivated participants, or whatever) may become frustrated and drop out, whereas in another condition it may be 5% of the lowest self-monitors (least motivated, etc.) who leave. The remaining subpopulations are not equivalent (despite having been randomly assigned at an earlier point) and hence may respond differently on the dependent measure. Demonstrations of no pretest or background differences between participants who quit and those who stay may provide some reassurance.

Social Comparisons among Conditions.

Cook and Campbell (1979, p. 56) listed “compensatory rivalry” and “demoralization” as potential issues when participants see treatments as differentially desirable. Participants may react to their perceptions of their good or bad outcomes by trying harder or slacking off, potentially producing differences on the dependent variable of interest. In many social psychological studies, effects of different conditions on participants’ mood represent another related potential confound. Mood may be influenced by many types of common manipulations and is known to affect a wide range of judgments and behaviors. For this reason, experimenters often assess participants’ mood and perform analyses designed to show that mood was not affected, or at least (when used as a covariate) that it cannot explain the effects obtained on the main dependent variable.

Other related issues are discussed by Abelson (1995). The entire issue of preventing potential confounds and artifacts in experiments is a large one (see Aronson et al., 1990; Miller, 1972; Rosenthal & Rosnow, 1969).

Designs for Studying Mediation

Experimental designs can be set up to test hypotheses about the mediation of causal relationships (Baron & Kenny, 1986; Judd, Yzerbyt, and Muller, Chapter 25 in this volume; MacKinnon, Fairchild, & Fritz, 2007). That is, the question is not just whether X causes Z , but whether X causes Y , which in turn causes Z . A typical design involves manipulating X and measuring both the dependent variable Z and one or more putative mediators Y . In this situation the traditional approach to obtaining statistical evidence for mediation is to estimate several separate regression equations and make specific comparisons as Baron and Kenny (1986) describe. Other, more modern approaches are more powerful while requiring fewer assumptions (see MacKinnon et al., 2007). However, all approaches where the mediator is measured suffer from a limitation imposed by the design: while experimental manipulation assures us that X causally precedes both Y and Z , there is no basis in the design for establishing causal priority between the latter two variables. The standard mediation analysis requires very strong assumptions: not only the obvious one that Y is causally prior to Z , but also that all unobserved causes of Y are uncorrelated with unobserved causes of Z (Bullock, Green, & Ha, 2010). For example, this assumption requires that X is the *only* common cause of both Y and Z , an assumption that often seems untenable. And these assumptions (because they involve unmeasured variables)

generally are not empirically testable, at least within a single study.

Based on considerations like these, experimental design approaches to mediation have been advocated (Bullock et al., 2010; Spencer, Zanna, & Fong, 2005). After conducting a study in which X is manipulated to establish its effects on both Z and Y , a second study manipulates Y and measures Z . This study allows for clear inferences that Y causes Z and thereby strengthens empirical support for the causal chain involved in mediation. However, even the experimental approach must make untestable assumptions. One is that the Y as measured in the first experiment is the same conceptual variable as the Y manipulated in the second experiment. A second is that the manipulations used in each of the experiments vary *only* the intended variables and not other potential mediators (including quite general constructs such as mood or general motivation to process information).

In the face of these difficulties, how can mediation be established? Current thinking (e.g., Bullock et al., 2010; Judd et al., Chapter 25 in this volume) is quite clear that mediation cannot generally be demonstrated with a single statistical procedure, or in a single study. At best, results from a study can be termed consistent with the conceptual mediation model (and researchers should devote attention to justifying the key assumptions, whether on the basis of theory or prior research). Instead, mediation can be firmly supported only by patterns of results from multiple studies, often with contributions from multiple research teams. For example, a series of studies might find that variants of the independent variable X that produce stronger effects on Z also tend to be those that have stronger effects on Y , while also showing that the Y - Z relationship is robust across such variations. Researchers are also turning increasingly to new types of evidence for hypotheses about mediating processes (beyond the statistical and design approaches described here): process dissociation procedures (Jacoby, 1991), response time or neuroscience measures that triangulate processes in other ways (Berkman, Cunningham, & Lieberman, Chapter 7 in this volume), or modeling procedures (where a model including a particular mediator is shown to fit the data, but competing models lacking that mediator fail to fit).

Concluding Comments

Design is basic to all research, in social psychology as well as other scientific fields. Yet in some ways classic treatments of design seem less than perfectly

applicable to current practice in social psychology and therefore offer incomplete guidance to a researcher struggling to make the best decisions about his or her work. Some of the ways in which this is true have been hinted at earlier (e.g., in the discussion of power analysis), but two related points will be made explicit now.

The background for both of these points is the observation that principles of research design were developed originally in the context of agricultural research and have been elaborated particularly by Donald Campbell and his colleagues (e.g., Cook & Campbell, 1979) as they thought about evaluations of educational innovations and large-scale social interventions. In all these areas, the key research question is ordinarily, “Does a given treatment have an effect?” and the goal is usually to answer the question with a single study. After all, a study may cost millions and take years (or at least an entire growing season!) to conduct. The influential Campbellian tradition of thinking about design has implicitly brought these emphases into social psychology. Yet our field often has questions that go beyond whether a treatment has an effect, and we are not restricted to answering our questions with large, single studies. Given these realities, what shifts in our thinking about research design are called for?

First, social psychologists today are interested in mediational questions – questions of how X causes Z – as often as, if not more often than, questions about whether X causes Z (see Brewer & Crano, Chapter 2 in this volume). Some earlier parts of this chapter describe ways design can help establish mediation. Yet more development is needed in this area. For example, most current treatments of power analysis focus on power to detect an effect but are less directly applicable to the issue of power to detect a mediational path. And current thinking about generalizability does not address issues of the generalizability of a mediational finding (as opposed to a simple causal effect).

Second and even more important, in social psychology today the most important unit of research is not the individual study. As Abelson (1995, p. 38) stated in a remark already quoted in this chapter, in social psychology, “real research investigations usually involve multiple studies conducted as part of an ongoing conceptual elaboration of a particular topic.” Such a series of conceptually related studies – a research program – is most often associated with a given investigator and his or her collaborators, but may include work from several distinct laboratories. The research program is not conveniently handled by the classic concepts of research design (power, generalizability, etc.), which focus on an individual study. Nor is it well captured by the metaanalysis

movement (e.g., Hedges & Olkin, 1985), which focuses on drawing empirical conclusions from large numbers of heterogeneous studies on a given issue. The program is larger than an individual study but too small a body of research for metaanalysis.

However, the research program is arguably the unit that is most important in the advancement of social psychology as a field, for several reasons. As Brewer and Crano (Chapter 2 in this volume) note, a single study can be almost definitively assumed to have good internal validity (based on its use of experimental design), but the broader forms of validity – construct and external – almost always emerge only from a series of interrelated studies that can compensate for each others' weaknesses. The conceptual replications inherent to a research program, besides strengthening construct and external validity, provide the best evidence supporting mediation (as described earlier). They also bring theoretical structure to a series of studies. Programmatically related studies generally focus on theoretical questions and build on one another rather than addressing superficially interesting but atheoretical questions in scattershot fashion. Perhaps for these reasons, isolated studies that are not part of a program seem to have relatively little scientific impact. Finally, the program is clearly the unit that is most relevant to the individual scientist trying to make decisions about how to conduct his or her ongoing research.

Decisions at the level of the research program really constitute strategy rather than tactics. I propose as a strategic maxim, that more, smaller studies are often better than fewer, larger ones – assuming they possess adequate power. One reason is that in a research program the primary questions of theoretical interest evolve over time. One large study may give a definitive answer to the question that one had two years ago (when that study was designed), but several smaller studies in the same amount of time will often lead to different questions. Conceptual advancement in research is measured as much by evolution in the questions that are asked as by the answers that are obtained. A second reason is that a series of relatively small studies is more likely to involve variation in nonessential factors (settings, measurement techniques, etc.), and therefore is likely to be stronger in construct and external validity, compared with a single massive study. This is particularly true if the studies are methodologically diverse rather than being variants on a single narrow paradigm (Reis & Gosling, 2010). The inevitable trade-off, of course, is power: A larger study always carries a higher probability of finding a significant effect.

A research program has all the same goals as an individual study – maximum

power, minimum cost, ability to rule out confounds, ability to generalize – and involves additional principles as well, such as Abelson's (1995) MAGIC criteria for the persuasiveness of data-based arguments, including “interestingness.” However, principles at this level have not yet been clearly codified so that they can be explicitly taught to apprentice researchers. Perhaps, just as Campbell's influential conceptualization of the forms of validity grew out of the context of massive studies evaluating large-scale social and educational interventions, an expanded treatment of research design that takes account of the research program level as well as the individual study may arise from the context of today's programmatic, mediationally focused research in social psychology.

References

- Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.
- Abelson, R. P. (1996). Vulnerability of contrast tests to simpler interpretations: An addendum to Rosnow and Rosenthal. *Psychological Science*, 7, 242–246.
- Aronson, E., Ellsworth, P. C., Carlsmith, J. M., & Gonzales, M. H. (1990). *Methods of research in social psychology* (2nd ed.). New York: McGraw-Hill.
- Bargh, J. A., Bond, R. N., Lombardi, W. J., & Tota, M. E. (1986). The additive nature of chronic and temporary sources of construct accessibility. *Journal of Personality and Social Psychology*, 50, 869–878.
- Baron, R. M., & Kenny, D. A. (1986). The mediator-moderator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Bem, D. J., Wallach, M. A., & Kogan, N. (1965). Group decision making under risk of aversive consequences. *Journal of Personality and Social Psychology*, 1, 453–460.
- Brown, R. (1986). *Social psychology* (2nd ed.). New York: Free Press.
- Brunswik, E. (1955). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: University of California Press.
- Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (don't expect an easy answer). *Journal of Personality and Social Psychology*, 98, 550--558.

- Bussemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, 93, 549–562.
- Bush, L. K., Hess, U., & Wolford, G. (1993). Transformations for within-subject designs: A Monte Carlo investigation. *Psychological Bulletin*, 113, 566–579.
- Campbell, D. T., & Boruch, R. F. (1975). Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experience: Some critical issues in assessing social programs* (pp. 195–296). New York: Academic Press.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145–153.
- Cohen, J. (1968). Multiple regression as a general dataanalytic system. *Psychological Bulletin*, 70, 426–443.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304–1312.
- Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin*, 114, 174–184.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation*. Chicago: Rand McNally.
- Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology*, 45, 545–580.
- Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, 95, 91–101.
- Feldt, L. S. (1958). A comparison of the precision of three experimental designs

- employing a concomitant variable. *Psychometrika*, 23, 335–354.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
- Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology*, 33, 175–183.
- Griffin, D., & Gonzalez, R. (1995). Correlational analysis of dyad-level data in the exchangeable case. *Psychological Bulletin*, 118, 430–439.
- Hastie, R., & Kumar, P. A. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality and Social Psychology*, 37, 25–38.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for metaanalysis*. New York: Academic Press.
- Higgins, E. T., King, G. A., & Mavin, G. H. (1982). Individual construct accessibility and subjective impressions and recall. *Journal of Personality and Social Psychology*, 43, 35–47.
- Hoyle, R. H., Harris, M. J., & Judd, C. M. (2002). *Research methods in social relations* (7th ed.). Belmont, CA: Wadsworth.
- Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin*, 117, 348–357.
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30, 513–541.
- Jöreskog, K. G., & Sörbom, D. (1991). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago: SPSS.
- Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions*. Cambridge: Cambridge University Press.
- Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-participant designs. *Psychological Methods*, 6, 115–134.

- Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model-comparison approach*. San Diego, CA: Harcourt Brace Jovanovich.
- Kenny, D. A. (1985). Quantitative methods for social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (3rd ed., Vol. 1, pp. 487–508). New York: Random House.
- Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, 99, 422–431.
- Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York: Guilford.
- Kenny, D. A., & Smith, E. R. (1980). A note on the analysis of designs in which subjects receive each stimulus only once. *Journal of Experimental Social Psychology*, 16, 497–507.
- Keppel, G. (1982). *Design and analysis: A researcher's handbook* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences*. Belmont, CA: Brooks/Cole.
- MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, 58, 593–614.
- Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, 113, 181–190.
- McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods*, 2, 3–19.
- McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114, 376–390.
- Miller, A. G. (1972). *The social psychology of psychological research*. New York: Free Press.
- Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, 38, 379–388.
- Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, 83, 602–627.

- Orne, M. (1962). On the social psychology of the psychological experiment. *American Psychologist*, 17, 776–783.
- Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology*, 60, 339–367.
- Petty, R. E., & Cacioppo, J. T. (1979). Issue involvement can increase or decrease persuasion by enhancing messenger-relevant cognitive responses. *Journal of Personality and Social Psychology*, 41, 847–855.
- Petty, R. E., Fabrigar, L. R., Wegener, D. T., & Priester, J. R. (1996). Understanding data when interactions are present or hypothesized. *Psychological Science*, 7, 247–252.
- Reis, H. T., & Gosling, S. D. (2010). Social psychological methods outside the laboratory. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., Vol. 1, pp. 82–114). New York: Wiley.
- Rock, D. A., Werts, C., & Flaughter, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of the multiple variables across populations. *Multivariate Behavioral Research*, 13, 403–418.
- Rosenthal, R., & Rosnow, R. L. (Eds.). (1969). *Artifact in behavioral research*. New York: Academic Press.
- Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge: Cambridge University Press.
- Rosnow, R. L., & Rosenthal, R. (1995). “Some things you learn aren't so”: Cohen's paradox, Asch's paradigm, and the interpretation of interaction. *Psychological Science*, 6, 3–9.
- Rosnow, R. L., & Rosenthal, R. (1996). Contrasts and interactions redux: Five easy pieces. *Psychological Science*, 7, 253–257.
- Santa, J. L., Miller, J. J., & Shaw, M. L. (1979). Using quasi *F* to prevent alpha inflation due to stimulus variation. *Psychological Bulletin*, 86, 37–46.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal influence*. New York:

Houghton Mifflin.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, 89, 845–851.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Zajonc, R. B. (1965). Social facilitation. *Science*, 149 (Whole No. 3681), 269–274.

* Thanks to Howard Weiss as well as the editors for helpful comments on the earlier versions of this chapter.

Chapter four Causal Inference and Generalization in Field Settings

Experimental and Quasi-Experimental Designs

Stephen G. West, Heining Cham and Yu Liu

The purpose of this chapter is to introduce researchers to randomized and nonrandomized designs that permit relatively strong causal inferences, particularly in field settings. We begin the chapter by considering some basic issues in inferring causality, initially drawing on work by Rubin and his associates in statistics (e.g., Holland, 1986, 1988; Imbens & Rubin, in press; Rubin, 1974, 1978, 2005) and later drawing on work by Campbell and his associates in psychology (Campbell, 1957; Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish & Cook, 2009; Shadish, Cook, & Campbell, 2002). Rubin's approach emphasizes formal statistical criteria for inference and the estimation of the precise magnitude of treatment effects; Campbell's approach emphasizes concepts from philosophy of science, practical issues confronting social researchers, and focuses more on the direction of effects (see Shadish, 2010; Shadish & Sullivan, 2012; West & Thoemmes, 2010). Rubin's approach offers statistical adjustments when problems occur; Campbell's approach offers methods for preventing problems from occurring. We initially apply these approaches to provide insights on difficult issues that arise in randomized experiments. Perspectives on the generalization of causal effects are also discussed (Shadish et al., 2002; Stuart, Cole, Bradshaw, & Leaf, 2011). We then consider three classes of quasi-experimental designs – the regression discontinuity design, the interrupted time series design, and the nonequivalent control group design – that can provide a relatively strong basis for causal inference when randomization is not possible. Rubin and Campbell's frameworks offer methods of strengthening each design type with respect to causal inference and generalization of causal effects.

The emphasis on designs for field research in the present chapter contrasts sharply with standard practice in basic social psychology (see Cialdini & Paluck, Chapter 5 in this volume, for a more general discussion of field research in social-personality psychology). The modal research design has long consisted of

a randomized experiment, conducted in the laboratory, lasting no more than one hour, and using undergraduate students as subjects (e.g., West, Newsom, & Fenaughty, 1992). This practice has clear strengths, notably in establishing internal validity – some component of the treatment caused the observed response (Brewer & Crano, Chapter 2 in this volume; Smith, Chapter 3 in this volume). It has a clear weakness in failing to provide a principled basis for generalizing the obtained results to the persons and contexts of interest. As complications are introduced into laboratory settings, claims of internal validity become more tenuous. Experiments conducted over repeated sessions may involve attrition. Participants may be suspicious of experimental procedures. Basic psychological processes that unfold over more extended time periods, that involve important societal interventions (changes in laws, introduction of new programs), or that require specialized contexts may be difficult to capture in the 50-minute laboratory session. As researchers move to field settings other challenging validity issues arise. The Rubin and Campbell frameworks provide principles for strengthening causal inference and generalization of research findings.

Rubin's Causal Model: One Framework for Causal Inference

Over the past three decades, Rubin and colleagues (e.g., Holland, 1986; Imbens & Rubin, in press; Rubin, 1974, 1978, 1986, 2005, 2011) have developed a framework known as the potential outcomes model (a.k.a. Rubin Causal Model, RCM) for understanding the causal effects of *treatments*. The RCM is particularly useful in identifying strengths and limitations of designs in which an independent variable is manipulated and posttest measures are collected at only one point in time following the treatment.

Consider the simple case of two treatments whose effects the researcher wishes to compare. For example, a basic researcher may wish to compare a severe frustration (treatment) and a mild frustration (comparison condition) in the level of aggressive responses they produce in participants (Berkowitz, 1993). Or, an applied researcher may wish to compare a 10-week smoking prevention program (treatment) and a no program group (control) on attitudes toward smoking among teenagers (Flay, 1986).

Rubin begins with a consideration of the *ideal* conditions under which a causal effect could be observed. These conditions *cannot* be realized in practice.

He defines the causal effect as the difference between what *would* have happened to a single participant under the treatment condition and what *would* have happened to the same participant in the control condition under identical circumstances. That is, the individual causal effect (CE) is defined as the difference between the participant's potential outcomes in the two conditions:

$$CE = Y_T(u) - Y_C(u)$$

Here, T refers to the treatment condition, C refers to the comparison condition (often a no-treatment control group), Y is the observed response (dependent measure), and u is the unit (typically a specific participant) on which we observe the effects of the two treatment conditions. Rubin's definition leads to a clear theoretical statement of what a causal effect is, but it also implies a fundamental problem. "It is impossible to observe the value of $Y_T(u)$ and $Y_C(u)$ on the same unit and, therefore, it is impossible to *observe* the effect of T on u " (Holland, 1986, p. 947, italics in original). We cannot expose a pre-teenage boy to the 10-week smoking prevention program, measure his attitudes toward smoking, then return the child to the beginning of the school year, expose him to the control program, and remeasure his attitudes toward smoking. Causality cannot be observed directly. By making specific assumptions, we can develop research designs that permit us to *infer* causality. The certainty of the causal inference will depend strongly on the viability of the assumptions. In the discussion that follows, we focus on the randomized experiment, one design approach that approximates Rubin's ideal conditions. The assumptions of this approach are often viable, so it is commonly used in both laboratory and field settings in social psychology.

Randomization as an Approach to the Fundamental Problem of Causal Inference

Rubin's ideal is to give the treatment and control conditions to the same participant at the same time and in the same context so the participant's potential outcomes in each treatment condition can be observed. Randomization approaches this ideal by approximately equating the treatment and control groups on all possible baseline covariates prior to any treatment delivery. Participants are assigned to treatment conditions using a method that gives every participant an equal chance of being assigned to the treatment and control conditions.¹ Common randomization methods include flipping a coin (e.g., heads = treatment; tails = control) or using a computer to generate random numbers

identifying treatment group assignments. Although randomization is normally straightforward in the laboratory, some field research settings pose very difficult randomization challenges. For example, in health care settings, participants often arrive when they become ill: The researcher must ensure that facilities and staff to deliver either the experimental or the control treatments are available at that time so randomization can be implemented. Boruch (1997) and Shadish *et al.* (2002) review methods that have been developed to address complex randomization issues.

Following random assignment, each participant then receives the treatment condition (e.g., experimental treatment vs. comparison [control] treatment) to which he or she was assigned. The responses of each participant are then measured after the receipt of the treatment or comparison condition. These procedures characterize the basic randomized experiment used in both laboratory and field settings in social psychology.

Random assignment means that the variable representing the treatment condition (treatment vs. control) can be expected on average to be *independent* of any measured or unmeasured variable prior to treatment. This outcome is reflected in three closely related ways.

1. Prior to any treatment delivery, the distributions of the treatment and control groups will, on average in large samples, be identical for any measured or unmeasured variable.
2. Prior to any treatment delivery, the expected values of the difference between the treatment and control groups, $E(\bar{Y}_T - \bar{Y}_C)$, will be zero for any variable Y . In other words, at pretest, the demographic characteristics, attitudes, motivations, personality traits, and abilities, can, on average in large samples, be expected to be the same in the treatment and control groups.
3. The treatment assignment variable X (where $X = T$ for treatment and C for control) will, on average in large samples, be unrelated ($r = 0$) to any measured or unmeasured participant variable prior to treatment.

These outcomes provide a new definition of the causal effect that is appropriate for randomized experiments. The *average causal effect* (ACE) at posttest is now defined as:

$$ACE = \bar{Y}_T - \bar{Y}_C$$

Three observations are in order. First, the comparison shifts to the ACE, the *average* response for the group of participants receiving the experimental treatment compared to the *average* response for the group of participants receiving the control treatment. Causal inferences about each individual may no longer be made. Second, the results are only expectations of what would occur “on average” in large samples. Exact equivalence of distributions and pretest means on pretest measures does *not* occur routinely in practice in any single sample. With simple randomization, exact pretest equivalence and exact independence of the treatment and background variables occur only if randomization is applied to a very large ($n \rightarrow \infty$) population or the results are averaged across a very large number (all possible) of different randomizations of the same sample. In any given real sample – that is, *your* sample – there is no guarantee that pretest means will *not* differ on important variables. If so, the estimate of the ACE can be too low or too high. Randomization replaces definitive statements about causal effects with probabilistic statements based on sampling theory. Third, for proper causal inference we need to make several assumptions discussed in more detail later in this chapter.

Finally, although randomization is normally performed using participants as the units, it may also be performed using other units. Reichardt (2006) provides a general framework and examples of how randomization can be applied to different classes of units. Most common is cluster randomization in which larger units such as laboratory groups, classrooms, or communities are randomized. Some social psychologists (e.g., Robert Cialdini) have randomly assigned different times or different locations to treatment and control conditions in their applied social research (West & Graziano, 2012).

Illustrative Example: Randomization

In Table 4.1 we present an example data set constructed based on Rubin's ideal case. There are 32 participants. The posttest response of each participant is observed under *both* the control (Y_C column) and treatment (Y_T column) conditions. The causal effect, $Y_T - Y_C$, is 0.5 for each participant. $\mu_C = 3.0$ and $\mu_T = 3.5$ are the means in the treatment and control groups, respectively; the standard deviation σ is 1.0. The distributions within each group are roughly normal. In this example, $d = \frac{\mu_T - \mu_C}{\sigma} = 0.5$, which represents a 0.5 standard deviation effect size, which Cohen (1988) describes as a moderate standardized effect size. As a benchmark, meta-analytic reviews of studies in both social psychology (Richard, Bond, & Stokes-Zoota, 2003) and personality (Fraley &

Marks, 2007) have found the average standardized effect size is $d = 0.21$. Table 4.1 also includes two columns labeled a_1 and a_2 . These illustrate two different possible random assignments of the 32 total participants to two equal groups ($n_T = 16$; $n_C = 16 = n$). There are $\frac{(2n)!}{n!n!}$ possible unique combinations of $2n$ participants with n in each group, which represents the number of different potential random assignments. Here, there are $\frac{32!}{16!16!}$, or more than 600 million different potential random assignments (Cochran & Cox, 1957).

The result of random assignment is that only one of the two potential outcomes can be observed for each participant. Two possible randomizations a_1 and a_2 are presented. Table 4.1 also illustrates the result of randomization a_1 . For participant 1, the response $Y_C = 1.0$ is observed under the control condition, but that the response $Y_T = \blacksquare$ is *not* observed in the treatment condition; for participant 2, the response is observed under the treatment, but not the control condition; and so on. The black squares represent the unobserved response for each participant.

Table 4.1. Illustration: Estimating the Causal Effect – Ideal Case and Randomized Experiment (RE) a_1

Participant	a_1	a_2	Ideal Case		RE a_1	
			Y_C	Y_T	Y_C	Y_T
1	0	1	1	1.5	1	■
2	1	0	2	2.5	■	2.5
3	1	1	2	2.5	■	2.5
4	1	0	2	2.5	■	2.5
5	0	0	2	2.5	2	■
6	0	1	3	3.5	3	■
7	1	1	3	3.5	■	3.5
8	1	0	3	3.5	■	3.5
9	0	1	3	3.5	3	■
10	0	1	3	3.5	3	■
11	0	1	3	3.5	3	■
12	1	0	4	4.5	■	4.5
13	0	0	4	4.5	4	■
14	0	1	4	4.5	4	■
15	1	1	4	4.5	■	4.5
16	0	0	5	5.5	5	■
17	1	1	1	1.5	■	1.5
18	1	1	2	2.5	■	2.5
19	0	0	2	2.5	2	■
20	1	0	2	2.5	■	2.5
21	1	1	2	2.5	■	2.5
22	0	1	3	3.5	3	■
23	1	1	3	3.5	■	3.5
24	1	1	3	3.5	■	3.5
25	0	0	3	3.5	3	■
26	0	0	3	3.5	3	■
27	1	1	3	3.5	■	3.5
28	1	0	4	4.5	■	4.5
29	1	0	4	4.5	■	4.5
30	0	0	4	4.5	4	■
31	0	0	4	4.5	4	■
32	0	0	5	5.5	5	■

Note: The column labeled Y_C contains the true response of each participant in the control condition. The column labeled Y_T contains the true response of each participant in the treatment condition. a_1 and a_2 represent two different random assignments of 16 participants to the control group and 16

participants to the treatment group. In each random assignment, 0 means the participant was assigned to the control group and 1 means the participant was assigned to the treatment group. ■ means response was not observed. In the ideal case, the mean of all 32 participants under the control condition is 3.0, the mean of all 32 participants under the treatment condition is 3.5, and the standard deviation of each condition is 1.0. The distribution within each group is approximately normal. The causal effect for each participant is 0.5, corresponding to a moderate effect size.

Figure 4.1 presents the distribution of the standardized effect sizes d calculated for each of the more than 600 million possible randomizations. Half of the estimates are less than $d = 0.5$ and about 8% are less than 0.² Only about 30% of the estimates (i.e., those $> +0.69$) would lead to correct rejection of the null hypothesis of no causal effect. This value – the probability of rejecting the null hypothesis when it is false – is known as the *statistical power* of the test. User-friendly software (e.g., G*Power 3; Faul, Erdfelder, Lang, & Buchner, 2007) can be used when designing the experiment to find out if the sample size is sufficient to detect the expected effect. Here, 64 participants in each treatment group (128 total) would be needed to detect the $d = 0.5$ effect with .80 power.

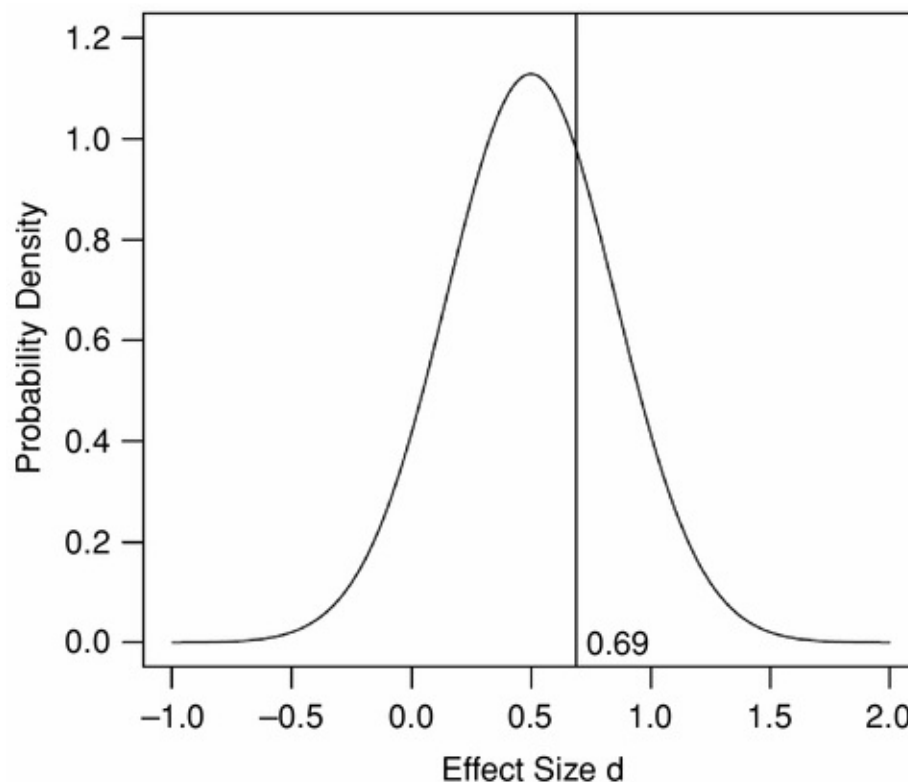


Figure 4.1. Sampling distribution of effect size d .

Note: The mean of the sampling distribution is 0.5. The standard deviation of the sampling distribution is 0.354. $\hat{Y}_T - \hat{Y}_C$ must exceed $1.96 \times 0.354 = 0.69$ to reject the null hypothesis of no difference between μ_T and μ_C in the population.

The experimenter will correctly reject the false null hypothesis about 30% of the time (statistical power).

Assumptions: Problems and Remedies

For randomization to lead to unbiased estimates of the ACE, a number of key assumptions must be met. These assumptions will often be met in laboratory experiments; they can be a challenge in field experiments. We present each of the assumptions in turn followed by potential design and analysis strategies that may be employed to prevent or remedy violations.

1. *Randomization was properly carried out.* For the benefits of randomization to accrue, random assignment of participants to treatment and comparison conditions must be properly carried out. Studies of large-scale randomized field experiments, particularly those conducted at multiple sites, suggest that full or partial breakdowns of randomization occur with some frequency (Boruch, McSweeny, & Soderstrom, 1978; Conner, 1977). Problems tend to occur more frequently when the individuals responsible for delivering the treatment, such as school or medical personnel, are allowed to carry out the assignment of participants to treatment conditions and when monitoring of the maintenance of the treatment assignment is poor. For example, Kopans (1994) reviewed a large Canadian experiment on the effectiveness of mammography screening for reducing deaths from breast cancer. He presented data suggesting that women in the mammography group had a substantially higher cancer risk *at pretest* than women in the no mammography screening group. Some physicians apparently assigned patients with known family histories or prior episodes of breast-related disease to the screening group.

This problem may be addressed through researchers maintaining control of the randomization process combined with careful monitoring of the treatment each participant actually receives following randomization (Braucht & Reichardt, 1993). For example, students (or their parents) in school-based experiments are sometimes able to agitate successfully to change from a control to a treatment class during the randomization process itself or during the school year following randomization. Careful monitoring can help minimize this problem and can also allow formal assessment of the magnitude of the problem. If there is a systematic movement of children between treatment conditions (e.g., the brighter children are moved to the treatment group), attempts need to be made to correct for the potential bias.

A design strategy for minimizing breakdowns of randomization is to use units that are temporally or geographically isolated in the experiment. Randomization breakdowns in school-based experiments are far more likely when different treatments are given to different classrooms (low isolation of units) than when different treatments are given to different schools (high isolation of units). King, Nielsen, Coberley, Pope, and Wells (2011) review many of the methods that can be used to facilitate the success of randomization.

2. *All participants were measured at posttest (no attrition)*. Participants cannot always be measured at posttest. Participants may move to another location, refuse to participate, or fail to complete key outcome measures. This problem of participant *attrition* can often be minimized by careful attention during the planning of the experiment. Securing the addresses and telephone numbers of the participants, their close friends or relatives, their employers or schools greatly aids in keeping participants who relocate in the experiment. Keeping in touch with both treatment and control participants through periodic mailings, telephone calls, or electronic communications and providing incentives for continued participation can also help minimize participant loss. Ribisl *et al.* (1996) present a summary of traditional techniques for tracking, locating, and contacting participants; newer, Internet-based techniques are constantly being developed. Nonetheless, even given the use of careful procedures, attrition still typically occurs. Biglan *et al.* (1991) in their review of longitudinal studies of substance abuse prevention reported attrition rates ranging from 5% to 66% (mean = approximately 25%). Furthermore, dropouts typically reported greater substance use at the initial measurement. Such findings suggest that estimates of treatment effects may be biased if attrition is not addressed.

When attrition does occur, modern methods of missing data analysis (Enders, 2010; Graham, 2012; Little & Rubin, 2002) should be used to provide less biased estimates of causal effects. If all systematic reasons for missing data are represented by other measured variables in the data set (data are termed *missing at random*³), then the estimate of the ACE can be properly adjusted, yielding an unbiased estimate of the causal effect. For example, if some asthma patients do not respond to an asthma symptoms questionnaire because it is scheduled during family visit times, inclusion of family visit times as a predictor in the treatment effect model can yield a properly adjusted estimate of the ACE. In contrast, if those patients fail to complete the symptoms questionnaire because they are having severe asthma symptoms (data are *missing not at random*), estimates of the ACE will be biased. If variables exist in the data set that are related to both

whether the data are missing (or not) and to the participant's reported or unreported posttest score on symptoms (e.g., baseline symptoms measure; breathing measures), the bias owing to missing data in the estimate of the ACE can be reduced, sometimes substantially.

Two modern missing data techniques that provide proper adjustment of data that are missing at random are full information maximum likelihood (FIML) estimation and multiple imputation (MI, Enders, 2010). FIML uses all information from the available data, including cases with partially missing data, to estimate the ACE. MI produces multiple copies of the data set (e.g., 20) in which the observed values are maintained and missing values are imputed from the available data using a sophisticated variant of regression analysis. MI retains the prediction error observed in the original data; each copy of the data set will have a different value ($\hat{y} + \text{error}$) imputed for the missing values. MI then recombines the multiple data sets to produce a single mean estimate of each of the parameters with appropriate standard errors. The two techniques produce comparable results in large samples if the same set of variables is included in the model. A key advantage of MI is that it is easier to base the imputation on many variables in a rich data set, so that it can be easier to approximate missing at random data and produce minimally biased estimates of the ACE. Commercial (e.g., *Mplus*, SAS) and noncommercial (e.g., NORM) computer software for FIML and MI estimation has become widely available. Techniques for addressing missing not at random data have been proposed (Enders, 2011), but they require very strong assumptions. When these assumptions are violated, these techniques typically lead to more biased estimates of the ACE than FIML or MI.

3. *All participants receive the full treatment to which they were assigned (Treatment Adherence).* When participants are randomly assigned to treatment and control conditions in field experiments, adherence to treatment assignment may be less than complete. There are two common variants. First, participants in the *T* condition may refuse the treatment, participants in the *C* condition may seek out the treatment, or both. Approximately one-third of the women in the treatment group of randomized trial of the effectiveness of screening mammograms were never screened, whereas a much smaller percentage of the women in the control group got mammograms outside the study (Baker, 1998). Second, participants in either the *T* or *C* conditions or both may only complete a portion of a multisession treatment program. In a parenting program for low-income Mexican-American families (Gonzales et al., 2012), families attended a mean of approximately five of the nine sessions that comprised the full

treatment.

Practical methods exist for minimizing nonadherence, notably making the program attractive to participants, removing barriers to program attendance (e.g., providing transportation or child care), giving participants incentives for program attendance, and only allowing those participants who are willing to participate in both the treatment and control programs to be randomized (Shadish et al., 2002). Despite these efforts, some participants assigned to the treatment condition never receive any treatment. We assume that the researcher was able to measure the dependent variable on all participants, including those who do not receive treatment.

Three statistical approaches have been typically taken to the first variant of the treatment nonadherence problem (termed treatment noncompliance in the statistical literature; see Sagarin, West, Ratnikov, Homan, Ritchie, & Hansen, in press; West & Sagarin, 2000 for reviews). *Intention to treat* analysis (ITT; Lee, Ellenberg, Hirtz, & Nelson, 1991) follows Sir Ronald Fisher's maxim of "analyze them as you've randomized them" (cited in Boruch, 1997, p. 199), comparing the mean response of all participants assigned to the treatment condition (regardless of whether or not they received treatment) with the mean response of all participants assigned to the control condition. This analysis typically yields conservative estimates of the causal effect so long as there are no missing data. It requires no assumptions beyond those required for the randomized experiment. *Analysis by treatment received* throws out all participants assigned to the treatment group who do not in fact receive treatment. Such a comparison will yield a *biased* estimate of the causal effect (with the direction of bias being unknown) unless the stringent assumption can be made that the participants who drop out of the treatment condition represent a random sample of the participants in that condition (missing completely at random), which is rarely the case.⁴ The *local average treatment effect* (LATE) compares the mean of the participants in the treatment group who actually received the treatment with an adjusted mean of participants in the control group who *would have received the treatment if offered* (Angrist, Imbens, & Rubin, 1996). Both the first and the third approaches potentially produce meaningful estimates of treatment effects; however, the ITT analysis estimates the causal effect of treatment assignment in the entire sample, whereas LATE estimates the causal effect of treatment only for those participants who actually would receive the treatment.

To understand these three approaches, consider the data presented in [Table](#)

4.2. The data from Table 4.1 have been reordered so that participants 1–16 are in the control group and participants 17–32 are in the treatment group. To simplify the example, participant 1 is identical to participant 17, participant 2 is identical to participant 18, and so on. Second, we have indicated a systematic pattern of noncompliance. In column c_1 , the five participants with the lowest scores prior to treatment do not accept the treatment.

Table 4.2. *Illustration of Effects of Treatment Noncompliance*

Participant	a_3	c_1	Y_C	Y_T
1	0	0	1 *	1.5
2	0	0	2 *	2.5
3	0	0	2 *	2.5
4	0	0	2 *	2.5
5	0	0	2 *	2.5
6	0	1	3 *	3.5
7	0	1	3 *	3.5
8	0	1	3 *	3.5
9	0	1	3 *	3.5
10	0	1	3 *	3.5
11	0	1	3 *	3.5
12	0	1	4 *	4.5
13	0	1	4 *	4.5
14	0	1	4 *	4.5

15	0	1	4 *	4.5
16	0	1	5 *	5.5
17	1	0	1 *	1.5
18	1	0	2 *	2.5
19	1	0	2 *	2.5
20	1	0	2 *	2.5
21	1	0	2 *	2.5
22	1	1	3	3.5 *
23	1	1	3	3.5 *
24	1	1	3	3.5 *
25	1	1	3	3.5 *
26	1	1	3	3.5 *
27	1	1	3	3.5 *
28	1	1	4	4.5 *
29	1	1	4	4.5 *
30	1	1	4	4.5 *
31	1	1	4	4.5 *
32	1	1	5	5.5 *

Note: The column labeled Y_C contains the true response of each participant in the control condition. The column labeled Y_T contains the true response of each participant in the treatment condition. a_3

represents the assignment of the first 16 participants to the control group and the second 16 participants to the treatment group. $c_1 = 1$ means participant follows the treatment or control condition as assigned. $c_1 = 0$ means participant is a never taker and does not comply when in the treatment condition. The starred value of Y is the value actually observed for each participant. As before, the true causal effect, $Y_T - Y_C$, is 0.5.

ITT analysis compares the observed data (indicated with an asterisk in [Table 4.2](#)) for participants assigned to the treatment with the observed data for participants assigned to the control group. The mean for the control group is as before $\bar{Y}_C = 3.0$. However, treatment group participants 17–21 did not adhere and received no benefit from treatment, correspondingly reducing the treatment group mean, $\bar{Y}_T = 3.344$. The causal effect estimate $\bar{Y}_T - \bar{Y}_C = 0.344$ – more than a 30% reduction in the effect size from the true value for compliers of 0.50.

Analysis by treatment received eliminates participants 17–21 who did not receive the treatment from the analysis. \bar{Y}_T is now $\frac{\sum_{i=22}^{i=32} Y}{11} = 4.045$. Thus, the causal effect estimate is $\bar{Y}_T - \bar{Y}_C = 4.045 - 3.000 = 1.045$, which in this case is considerably larger than the true value of 0.5.

The LATE approach, based on Rubin's potential outcomes perspective, highlights an easily overlooked point. Participants 1–5 in the control group are identical to the nonadherers in the treatment group (participants 17–21) in [Table 4.2](#). These five participants would *not* have complied with the treatment *if* they had been assigned to the treatment group. All standard analyses that throw out noncompliers fail to take into account this group of participants who would fail to take the treatment if they were given the opportunity. Angrist *et al.* (1996) term these participants *never takers* as they would never agree to receive the treatment (whether they are assigned to the treatment or control groups). We eliminate the never takers from *both* the treatment and control groups with the result that \bar{Y}_T is now $\frac{\sum_{i=22}^{i=32} Y}{11} = 4.045$ and \bar{Y}_C is now $\frac{\sum_{i=6}^{i=16} Y}{11} = 3.545$, so that the LATE estimate of the causal effect is 0.5. This equals the true causal effect, but only for those participants who would accept the treatment if offered and the control if offered, termed *compliers* (or *adherers*).

Given a randomized experiment, the Mplus program can be used to calculate the LATE estimate and its standard error (Jo, 2002). This analysis assumes that the effect of treatment assignment only operates through the active treatment (e.g., there are no expectancy or placebo effects). This assumption is often, but not always, reasonable (see Hirano, Imbens, Rubin, & Zhou, 2000 for an

exception); masking treatment providers and participants to treatment assignment can eliminate the possibility that this assumption is violated. The LATE estimate is unbiased whereas the ITT estimate is typically attenuated (too close to 0). When both treatment nonadherence and attrition occur, the ITT estimate may potentially even be in the wrong direction (Frangakis & Rubin, 1999; Hirano et al., 2000).

The LATE estimate is available only in designs in which the control group represents “no treatment” or in which the control group represents a base treatment (T_{Base} that everyone receives) to which one or more additional components are added in an enhanced treatment group ($T_{Base} + T_{Additional}$; West & Aiken, 1997). Designs in which an alternative treatment is used as the comparison group will not produce a proper LATE estimate. Procedures also exist for estimating causal effects in the second variant described earlier in which some participants adhere to only part of the full treatment or control regimen. Given careful measurement of dosage received, a causal effect can be estimated comparing participants in the treatment group who receive a specified dosage with participants in the control group who would receive the same dosage if treatment were offered (Efron & Feldman, 1991; Holland, 1988; Jo, Ginexi, & Ialongo, 2010; see Sagarin et al., in press, for a review).

One drawback of the LATE estimate (and similar procedures that provide proper adjustment for partial adherence) is that they have a large standard error and therefore low statistical power because of uncertainty in making the proper adjustment for the never takers in the control condition who are not observed. If baseline covariates are available that predict adherence with treatment assignment within the treatment group, statistical power can potentially be dramatically increased (Jo, 2002).

4. *The Stable-Unit-Treatment-Value Assumption (SUTVA).* Rubin's perspective makes stipulations related to (a) the independence of the units and (b) the constancy of the delivery of the treatment and control conditions across participants. With respect to (a), it requires that there is a single potential outcome in the treatment condition and a single potential outcome in the control condition for each participant. In some cases, the outcome of the participant may depend on the treatment assignment of another participant. For example, sexually active participants in a randomized experiment evaluating a university safer sex program may have a lower risk of acquiring a sexually transmitted infection if their primary sex partner is also assigned to the treatment rather than the control program. In such a case there are two potential outcomes for each

treatment group participant depending on the partner's treatment assignment. Geographical or temporal isolation of participants can minimize this issue. Statistical correction methods for interference between units are presented in Rosenbaum (2007) and Sobel (2006).

With respect to (b), violations of SUTVA may also occur if there are hidden variants of treatment, as can occur when the treatment is implemented at multiple treatment sites or by multiple program staffers (e.g., experimenters). Some staff may effectively deliver the program producing a large treatment effect, whereas others' delivery may be far less effective so that their treatment effect is small. Replications of the experiment will be problematic because the estimate of the causal effect will depend on the peculiar mix of effective and ineffective program staff involved in the particular experiment. Training treatment providers to a criterion for delivering the treatment can prevent this problem. Measuring the quality of treatment delivery and including it in the statistical analysis can potentially adjust for treatment implementation problems (Sechrest, West, Phillips, Redner, & Yeaton, 1979). Alternatively, given sufficient data, staff or site effects can be modeled using dummy variables or as random effects (given at least 20 staff members; Kreft & de Leeuw, 1998) using multilevel models (Hox, 2010; Raudenbush & Bryk, 2002; Schoemann, Rhemtulla, & Little, Chapter 21 in this volume; Willett & Singer, 2013).

Group Administration of Treatment

In clustered randomized designs, interventions are delivered to groups of participants (Donner & Klar, 2000; Murray, 1998; Murray, Varnell, & Blitstein, 2004). Aiken, West, Woodward, and Reno (1994) delivered a treatment program encouraging compliance with American Cancer Society guidelines for screening mammograms or a no treatment control program to women's groups in the community. Group administration of treatments, whether in the laboratory or in the field, leads to statistical and inferential issues that need to be considered.

In clustered randomized designs, the entire group is assigned to receive either the treatment or control condition. The statistical outcome of this procedure is that the responses of the members of each treatment group may no longer be independent. In the Aiken *et al.* (1994) mammography experiment, the responses of two women randomly chosen from a single community group would be expected to be more similar than the responses of two women randomly chosen from different community groups. While nonindependence has no impact on the causal effect estimate, $\bar{Y}_T - \bar{Y}_C$, it does lead to estimates of the standard error

of this effect that are typically *too small*⁵ (Kenny & Kashy, Chapter 22 in this volume). The magnitude of this problem increases as the amount of dependence (measured by the intraclass correlation) and the size of groups to which treatment is delivered increase. Barcikowski (1981) showed that, even with relatively low levels of dependency in groups (intraclass correlation = .05), when groups were of a size typical of elementary school classrooms ($n = 25$ per class), the type 1 error rate (rejecting the null hypothesis when in fact it is true) was in fact .19 rather than the stated value of .05.

Following Fisher (1935), researchers traditionally “solved” this problem by aggregating their data to the level of the group (e.g., average response of each classroom). As mentioned earlier, the unit of analysis should match the unit of assignment – “analyze them as you’ve randomized them” (Fisher, cited in Boruch, 1997, p. 195). However, this solution is often not fully satisfactory because such analyses can be generalized only to a population of groups, not to a population of individuals, the latter being typically the researcher’s interest. Over the two past decades, new statistical procedures termed hierarchical linear (a.k.a. random coefficient, multilevel) models have been developed. These models simultaneously provide an estimate of the causal effect at the group level as well as individual-level analyses that appropriately correct standard errors for the degree of nonindependence within groups. Introductions to these models are presented in Schoemann, Rhemtulla, and Little (Chapter 21 in this volume), Snijders and Bosker (2011), and West, Ryu, Kwok, and Cham (2011).

Given the proper use of hierarchical linear models to analyze the data, random assignment of groups to treatments equates the groups at baseline on group-level variables. However, Rubin’s causal model emphasizes that causal effects represent the comparison of one well-articulated treatment with another well-articulated treatment. In comparisons of treatments delivered in group settings, the inference about the causal effect becomes murkier: Each participant’s potential outcomes may depend on the specific cluster to which they are assigned. Hong and Raudenbush (2006) and Thoemmes and West (2011) discuss some of the issues, distinguishing between cases in which clustering is an incidental feature of the design (e.g., participants are randomly assigned to clusters that are then randomly assigned to treatments) versus cases in which intact groups are assigned to treatments.

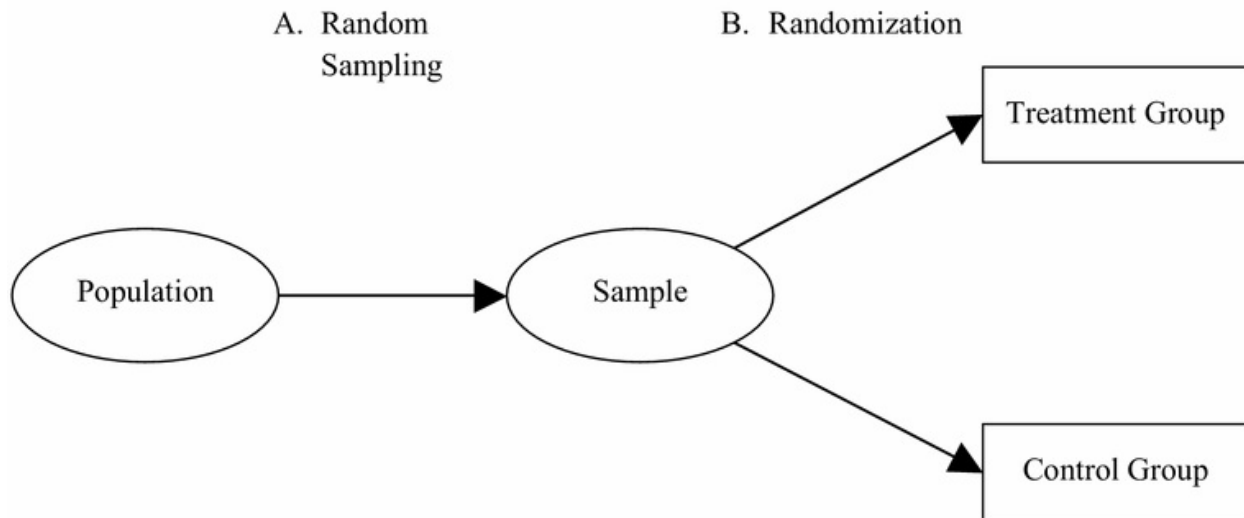


Figure 4.2. The formal statistical model for generalization.

Note: The purpose of Stage B is to provide unbiased estimates of the causal effect of the treatment. The purpose of Stage A is to permit generalization of the results obtained in the sample to a defined population of participants.

Generalization of Causal Relationships

So far our presentation has emphasized the Rubin Causal Model, which focuses primarily on the estimation of the causal effect, what Campbell (1957) originally termed *internal validity*: Something about the particular treatment caused the specific observed response in this particular context. Campbell (1986) later emphasized the limited causal understanding internal validity provided, even suggesting that internal validity be relabeled as “local molar (pragmatic, atheoretical) causal validity” (p. 69).

Social psychological researchers give high priority to internal validity, but they are rarely interested in limiting their causal statements to a specific treatment implementation, delivered in a specific setting, to a specific sample of participants, assessed with a specific measure. They wish to be able to generalize their causal statements. Cronbach (1982) has developed a useful framework for generalization with respect to four dimensions: units (participants), treatments, observations (outcome measures), and settings. He labels the values of these dimensions that characterize a specific experiment as (lower case) **units**, **treatments**, **observations**, and **settings** (utos). He labels the target dimensions to which the researcher can generalize the results of the experiment as (upper case) **Units**, **Treatments**, **Observations**, and **Settings** (UTOS).

Consider a classic experiment by Evans *et al.* (1981), which investigated the

effects of a smoking prevention program versus a no treatment control. In this experiment, the following utos were realized: the units were children in specific schools in Houston, Texas, the treatment was this specific implementation of his social influences smoking prevention program, the observations were children's self reports of smoking, and the setting was the classroom. The UTOS to which Evans and colleagues presumably wished to generalize were all adolescent school children, the specific social influences smoking prevention program, actual cigarette smoking, and school classrooms. As a focused applied experiment, the goal is to generalize to a specific population and to a specific school setting. The treatment is conceptually quite circumscribed as is the observation of interest – smoking.

Strategies for Generalization

Statistical Strategy: Sampling from a Defined Population. The only formal statistical basis for generalization is through the use of random sampling from a well-defined population. Surveys using national probability samples assess the attitudes of representative samples of adults; other surveys may assess the attitudes of a more focused sample, such as hospital nurses in a section of Denver, Colorado. In each case, a defined population is enumerated from which a random sample is collected, yielding estimates of the attitudes of the population that are accurate within a specified level of sampling error.

Figure 4.2 presents the randomized experiment in the context of this formal sampling model. Stage A represents random sampling from a defined population; the purpose of this stage is to assure generalization to a defined population of Units (participants). Stage B represents random assignment of the units in the sample to treatment and control conditions; as discussed previously, the purpose is to achieve unbiased estimates of the causal effect in the sample. The combination of Stages A and B formally permits generalization of an unbiased causal effect of the specific treatment conditions, studied in the context of a specific experimental setting, using the specific dependent measures to the full population of Units (participants). Note that generalization to Treatments, Observations, and Settings is *not* formally addressed by this model.

This formal sampling model is routinely recommended by statisticians and epidemiologists (e.g., Draper, 1995; Kish, 1987) as the ideal model for experimental research. Unfortunately, it is extraordinarily difficult to implement in practice. With respect to Units (participants), many populations cannot be precisely enumerated (e.g., children whose parents are alcoholics; individuals in

an Internet experiment). Even when the researcher can precisely enumerate the participant population (e.g., using official court records to enumerate recently divorced individuals), there is no guarantee that such participants can actually be located. Even if they are located, participants may refuse to participate in the experiment, despite being offered large incentives for their participation. Loss of participants can substantially restrict the generalization of results.

A few experiments have approximated the ideal model of statisticians. Randomized experiments have been conducted within national or local probability surveys (e.g., investigating question context; Schwarz & Hippler, 1995). Randomized experiments have compared treatment versus control programs using random samples of specific populations of individuals in the community (e.g., job seekers selected from state unemployment lines; Vinokur, Price, & Caplan, 1991). Such experiments routinely include heroic efforts to study nonparticipants in the experiment in order to understand the probable limits, if any, on the generalization of the findings to the full participant population of interest. If both the population and the achieved sample have been assessed on key background variables, weighting methods can be used to estimate the causal effect that would have been obtained if all sampled individuals had participated (Lohr, 2010; Stuart et al., 2011). However, such efforts only address the generalization of *units* from *u* to *U*. In laboratory experiments, stimuli (e.g., stimulus persons) are also sometimes randomly sampled from a population, permitting generalization when appropriate statistical models are used (Judd, Westfall, & Kenny, 2012). When our interest turns to the generalization of findings to a population of Treatments, Observations, and Settings, we almost never have any strong basis for defining a population of interest. In nearly all basic and applied social psychological research, we have to turn to extra-statistical methods of enhancing causal generalization.

Extra-Statistical Approaches: Five Principles. Cook (1993) and Shadish et al. (2002) synthesized earlier ideas and articulated five general principles supporting causal generalization. These principles may be applied to strengthen causal generalization with respect to Units, Treatments, Observations, and Settings.

1. *Proximal Similarity.* The specific units, treatments, observations, and settings should include most of the components of the construct or population, particularly those judged to be prototypical. A researcher wishing to generalize to a population of nurses (e.g., in metropolitan Denver) should choose nurses

from this area in his sample. The sample should include the various modal types of nurses (e.g., LPN, RN). To generalize to settings, the modal settings in which nurses work (e.g., hospital, home-care) should be identified and nurses sampled from each. With respect to constructs, the researcher should select or design a treatment and a measurement instrument that includes the key components specified by the theory.

2. *Heterogeneous Irrelevancies*. The units, treatments, observations, and settings used in our experiments are specific instances chosen to represent the population or the construct of interest. They will typically underrepresent certain features of the target Units, Treatments, Observations, and Settings. They will typically also include other extraneous features that are not part of the target of generalization. For example, nearly all research on attitudes uses paper and pencil or similar techniques adapted to a computer-based format, yet these measurement techniques are not part of the definition of attitudes (Henry, 2008; Houts, Cook, & Shadish, 1986; Sears, 1986). The principle of heterogeneous irrelevancies calls for the use of multiple instances in research, which are heterogeneous with respect to aspects of units, treatments, observations, and settings that are theoretically expected to be *irrelevant* to the treatment-outcome relationship. To the degree that the results of the experiment are consistent across different types of Units, different manipulations of the independent variable (Treatments), different Observations, and different types of Settings, the researcher can conclude that generalization of the findings is not limited.

3. *Discriminant Validity*. Basic social psychological theory and theories of programs (MacKinnon, 2008; West & Aiken, 1997) specify the processes through which a treatment is expected to have an effect on the outcome. These theories identify specific Treatment constructs (causal agents) that are supposed to affect specific Observation constructs (dependent variables). For example, the specific Treatment of frustration, defined as the blockage of an ongoing goal-directed behavior, is hypothesized to lead to increases in aggression. Other similar treatments that do not involve goal blockage such as completing an exciting task should *not* produce aggression. Given the focus of the hypothesis on aggression, the researcher should be able to show that frustration does not lead to other emotion-related responses (e.g., depression). If (a) the causal agent of the Treatment (here, frustration) is shown to match the hypothesized construct and (b) the class of Observations (here, aggression) affected by the Treatment matches those specified by the theory, claims for understanding the causal relationship are strengthened. This same approach can be taken to Units and Settings when hypotheses identify specific classes of units or settings over which

the causal effect theoretically should and should not generalize.

4. *Causal Explanation.* To the extent that we can support a causal explanation of our findings and rule out competing explanations, the likelihood of generalization is increased. The causal explanation distinguishes the active from the inert components of our treatment package and provides an understanding of the processes underlying our phenomenon of interest (MacKinnon, 2008; West & Aiken, 1997). These features permit us to specify the components that need to be included in any new experimental context. This principle has long been at the heart of basic experimental work in social psychology with its focus on the articulation and ruling out of competing theoretical explanations (Smith, Chapter 3 in this volume; Wilson, Aronson, & Carlsmith, 2010). More recently, both basic and applied social psychologists have used mediation analysis (Baron & Kenny, 1986; Judd, Yzerbyt, & Muller, Chapter 25 in this volume; MacKinnon, 2008) as a means of probing whether their favored theoretical explanation is consistent with the data. To the extent that the data are shown to support the favored theoretical explanation, mediation analysis can provide strong guidance for the design, implementation, and evaluation of future programs. However, mediation analysis does not automatically rule out other competing explanations for the observed effects; it requires careful consideration of potential baseline variables that may confound the relationship between the mediator and the outcome (Imai, Keele, & Tingley, 2010; Mayer, Thoemmes, Rose, Steyer, & West, 2013; Steyer, Partchev, Kroehne, Nagengast, & Fiege, in press). To the extent that these alternative causal explanations (a) are plausible and (b) make different predictions when new units (participants), new treatments, new observations, and/or new settings are studied, generalization of the findings of an experiment will be limited.

5. *Empirical Interpolation and Extrapolation.* For ease of presentation, this chapter has focused on the comparison of two treatment conditions. Although experiments are conducted in social psychology with more than two levels of each separate treatment variable, such experiments are not common, particularly in applied social research. In general, a high dose of the treatment (e.g., a smoking prevention program carried out over three years; Evans et al., 1981) is compared with a no treatment (or minimal treatment) control group. This practice yields an estimate of the causal effect that is limited to the specific implementations of treatment and control groups used in the experiment. Occasionally, parametric experiments or dose response (response surface) experiments involving one or more dimensions (Box & Draper, 1987; Collins, Murphy, Nair, & Streicher, 2005; West, Aiken, & Todd, 1993) can be

conducted. Such experiments help establish the functional form of the relationship between the strength of each treatment variable and the outcome of interest. If the functional form of the dose response curve is known, causal effects for the comparison of any pair of treatment conditions can be estimated through interpolation.

In the absence of such dose-response curves, caution must be exercised in generalizing the magnitude of causal effects very far beyond the specific levels of the treatment and control groups implemented in a particular experiment. The functional form of dose-response relationships may be nonlinear. Complications like threshold effects, the creation of new processes (e.g., psychological reactance if an influence attempt becomes too strong), and the influence of interactions with other background variables become increasingly likely as the gap increases between the treatments studied and those to which the researcher wishes to generalize. If we extrapolate beyond the range of units, treatments, observations, or settings used in previous research, our generalization of estimates of treatment effects become increasingly untrustworthy.

Summary. Traditional social psychological perspectives have relied nearly exclusively on the single principle of causal explanation. Cook's five principles add other criteria beyond causal explanation that can help enhance the likelihood that the causal effect will generalize to the UTOS of interest. Matt (2003), Matt and Cook (2009), and Shadish *et al.* (2002) also note that application of the five principles to meta-analyses of entire research literatures can provide an even stronger basis for generalization. Shadish, Hu, Glaser, Knonacki, and Wong (1998) present an illustration of how a response surface can be constructed in the context of a meta-analysis to estimate a variety of treatment effects of interest.

Quasi-Experimental Designs

Randomized experiments are typically considered to be the gold standard for causal inference. However, when researchers wish to address important applied questions, randomized experiments may be infeasible, unethical, not accepted in the research context, or only atypical participants may be willing to be randomized. Experiments on important questions such as the effects of secondhand tobacco smoke or sustained prejudice on health cannot be undertaken. Alternative quasi-experimental designs in which participants are nonrandomly assigned to treatment and control conditions can permit relatively strong causal inferences, yet may be conducted with the populations, treatments,

outcome measures, and in the settings of interest, maximizing causal generalization. Complementing Rubin's causal model presented earlier, Campbell's perspective offers a framework for considering quasi-experimental designs that is particularly useful for strengthening and generalizing causal inferences.

Campbell's Perspective: A Second Approach to Causal Inference

Campbell's (1957; Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish et al., 2002) perspective is familiar to many psychologists and is complementary to the Rubin causal model. Campbell begins with the identification of *plausible* threats to internal validity, reasons why researchers may be partly or completely wrong in making a causal inference in a specific research context. Major threats to internal validity are presented in Table 4.3 (see also Campbell & Stanley, 1966; Shadish et al., 2002). The task for researchers is to identify those specific threats that may operate in their specific design and research context; not all threats will be plausible. Then, design elements are incorporated that address the identified threats. The original hypothesis and the alternative artifactual explanation associated with each threat to internal validity are expected to produce different patterns of results; the obtained results are compared to each of the expected patterns to determine which provides the best account of the data – a process termed *pattern matching*. For example, replication of a treatment effect at nonoverlapping points in time rules out the possibility that history accounts for the results. Measuring growth at several time points and showing that the growth rates do not differ *prior* to the introduction of treatment can help rule out the possibility that different rates of maturation account for the differences between the treatment and control groups. Following the presentation of each of the basic quasi-experimentation designs in the sections that follow, we illustrate how the addition of design elements can strengthen causal inferences. Detailed lists of potential design elements can be found in Shadish and Cook (1999) and Shadish *et al.* (2002).

Table 4.3. Major Threats to Internal Validity in Quasi-Experimental Designs

Threats Arising from Participants' Growth and Experience

1. History: An event that is not part of the treatment occurs between the pretest and posttest that may affect the outcome.
2. Maturation: Within-participant processes unrelated to the treatment such as natural growth or decline may affect the outcome.

Threats Associated with the Measurement Process

3. Testing: Scores on subsequent measures may be affected by prior measurement.
4. Instrumentation: Changes in the capabilities of the measuring instrument may occur across measurements. Aggressive acts at age 4 are different from aggressive acts at age 16. Record keeping practices may change. Floor or ceiling effects may occur.

Threats Associated with Sampling Design

5. Statistical Regression: Participants selected on the basis of extreme high or low scores at time 1 may score closer to the group mean at time 2 even in the absence of treatment.
6. Selection: Participants in treatment and control groups may not be comparable at baseline prior to Treatment.
7. Interactions with Selection: Any of the threats 1–5 can interact with selection to affect the outcome.
8. Attrition: Participants may fail to complete the outcome measurement.

Note: Campbell and Stanley (1966) and Shadish, Cook, and Campbell (2002) present fuller discussions of these threats. In what follows we discuss the operation of the relevant threats in the context of specific experimental and quasi-experimental designs. Only those threats that are plausible in the context of a specific design need to be addressed. The randomized experiment rules out threats 1–7, but differential attrition can occur in the treatment and control groups, potentially leading to bias in the estimation of the causal effect.

Regression Discontinuity Design

One of the strongest alternatives to the randomized experiment is the regression discontinuity (RD) design. The RD design can be used when treatments are assigned on the basis of a baseline quantitative measure, often a measure of need, merit, or risk. Following Reichardt (2006), we term this measure the *quantitative assignment variable*. To cite three examples, at some universities, entering freshmen who have a verbal SAT score below a specified value (e.g., 380) are assigned to take a remedial English course, whereas freshman who have a score above this value are assigned to a regular English class (Aiken, West, Schwalm, Carroll, & Hsiung, 1998). The outcome of interest is the students' performance on a test of writing skill. In economic downturns, union members with fewer than a specified number of years of seniority (e.g., fewer than 20 years) are laid off, whereas more senior workers are retained. Mark and Mellor (1991) compared the degree of hindsight bias (retrospective judgments of the perceived likelihood of layoffs) among laid-off workers and workers who survived layoffs. Head Start programs were initially assigned to the poorest counties in the United States, those that had more than 59.2% of families living below the poverty line. Ludwig and Miller (2007) found that these counties experienced long-term improvements in educational outcomes of their children relative to financially slightly better-off counties that had fewer than 59.2% of the families living below the poverty line who did not initially receive Head Start. The key feature of the basic (sharp) RD design is that units (participants) are assigned to treatment or control conditions solely on the basis of whether they exceed or are below a cut point on the quantitative assignment variable and that an outcome hypothesized to be affected by the treatment is measured following treatment. The RD design meets the objection to randomized experiments that potentially beneficial treatments should not be withheld from the neediest (or more deserving) participants.

Consider a study by Seaver and Quarton (1976) illustrated in Figure 4.3. These researchers were interested in testing the hypothesis that social recognition for achievement leads to higher levels of subsequent achievement. As at many universities, those students who earned a Fall quarter grade point average (GPA) of 3.5 or greater were given formal recognition by being placed on the Dean's list, whereas students who did not attain this GPA were not given any special recognition. The grades for all students were recorded at the end of the Winter quarter as a measure of subsequent achievement. For simplicity, we assume that there is a linear relationship between Fall and Winter quarter GPAs.

Given this assumption, a linear regression equation

$$\widehat{GPA}_{Winter} = b_0 + b_1 GPA_{Fall} + b_2 (Dean's List)$$

can be used to estimate the discontinuity between the segments of the regression line below and above $GPA_{Fall} = 3.5$. This value is represented by b_2 in the equation. Seaver and Quarton estimated that Dean's list led to a 0.17 point increase in the students' GPA, the equivalent of a full grade higher in one three-hour course.

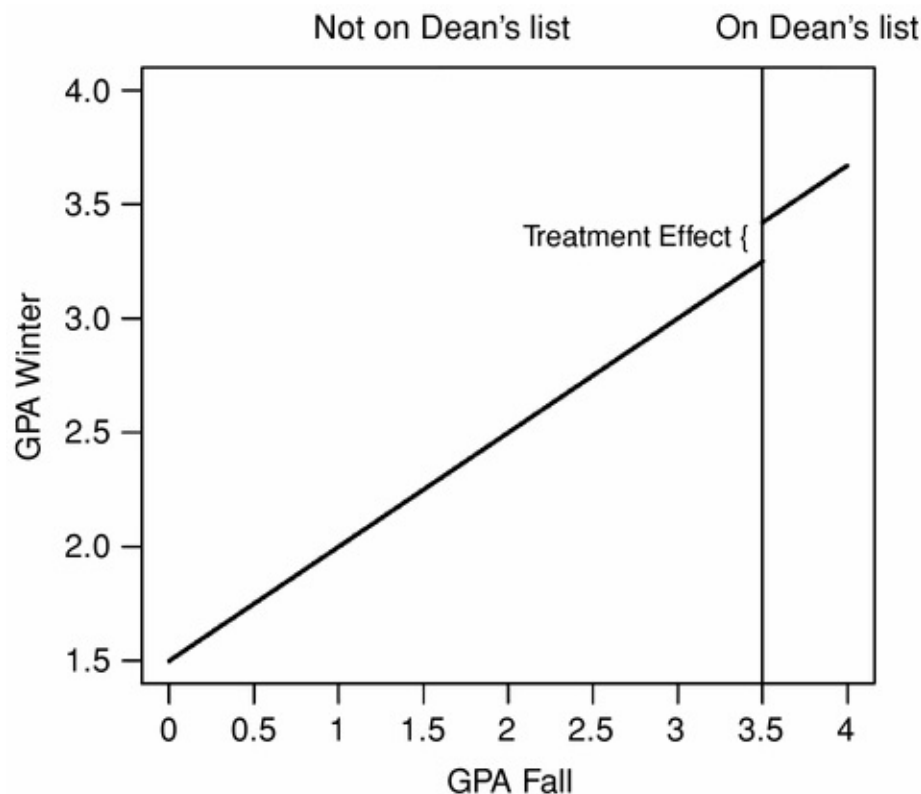


Figure 4.3. Regression of GPA Winter on GPA Fall for the non-dean's list and dean's list groups. *Note:* The vertical distance between the regression lines for the dean's list and for the control students (the discontinuity) represents the treatment effect (b_2).

From the perspective of Rubin's causal model, the RD design takes advantage of the known rule for assignment to treatment and control groups. The predicted value of the treatment group is compared with the predicted value of the control group at the cut point, conditioned on the specific value of the quantitative assignment variable ($GPA_{Fall} = 3.5$). So long as the quantitative assignment

variable is the sole basis on which participants are given the treatment versus control, this adjusted difference potentially provides an unbiased estimate of the treatment effect.

Verifying Adherence to the Assignment Rule.

Like the randomized experiment, the basic (sharp) RD design makes the key assumption that the assignment rule is rigorously followed. In our dean's list illustration, we assumed the university rigorously adhered to the 3.5 GPA_{Fall} cutoff rule, which is highly plausible. In other contexts, participants may know the cutoff in advance and may make a decision to opt into the study based on whether their score on the assignment variable would assign them to their preferred treatment condition. Or, practitioners who are responsible for assessing participants may alter scores on the assignment variable with the goal of assigning the participant to the treatment condition they deem optimal. Such failures of the assignment rule give rise to the possibility that selection accounts for the results. These problems can be minimized by not announcing the cutoff score until after participants are enrolled in the study. In addition, sophisticated statistical methods for probing adherence to the assignment rule have been developed in econometrics (Imbens & Lemieux, 2008). Here, we present three graphical methods for probing adherence to the assignment protocol.

1. Plot a graph of the assignment variable on the X-axis versus the binary treatment assignment ($T = 1$, $C = 0$) on the Y-axis. For the sharp RD design, there should be a vertical jump from $Y = 0$ to $Y = 1$ at the cut point.
2. Plot graphs of other covariates on the X-axis versus the binary treatment assignment. There should not be any discontinuity in level at the cut point.
3. Plot graphs of the distribution of the assignment variable. The distribution should be regular near the cut point. If there is an irregularity in the distribution, it may represent alteration of the scores on the assignment variable. For example, classroom teachers who score a qualification test for a state scholarship might add a few points to the score of prized students to make those students eligible for the scholarship (treatment), leading to irregularities in the distribution near the cut point.

Correct Specification of Functional Form.

In our dean's list example, we assumed that the relationship between Fall and Winter GPA was linear and that the effect of dean's list recognition would be to increase the subjects' GPA. To the degree that the form of the relationship between Fall quarter and Winter quarter GPA in our example is nonlinear, the estimate of the treatment effect based on the above regression equation may be biased. Two simple probes of the functional form may be performed.

First, scatterplots of the relationship between the quantitative assignment variable (e.g., Fall quarter GPA in our earlier example) and (a) the outcome measure and (b) the residuals should be carefully examined. Of particular importance in context of the RD design is the existence of systematic positive or negative residuals from the regression line near the cut point, suggesting a potentially major problem in the estimation of the treatment effect. This examination is greatly facilitated by the use of modern graphical techniques that fit nonparametric curves that describe the functional form in the data (e.g., lowess curves, Cleveland, 1993; smoothing splines, Green & Silverman, 1994). Second, the regression equation may be estimated using a nonlinear functional form. The simplest method is to add higher-order terms to the regression equation if there appears to be curvilinearity (e.g., X^2) or a change in slope (e.g., $X \times T$ interaction) at the cut point. These added terms may be tested for statistical significance. Alternatively, given large sample size, nonparametric regression methods, particularly methods developed specifically for the RD design proposed by Hahn, Todd, and Van der Klaauw (2001), may be used. If similar estimates of the treatment effect are found, the possibility that an incorrect functional form accounts for the results can be minimized.

When significant changes in slope are detected at the cut point, researchers need to be cautious in their interpretation of the results. If there is no discontinuity between the two regression lines (i.e., no treatment main effect), differences in slope are *not* usually interpretable. Given a change in slope, the absence of discontinuity increases the plausibility that there is not a treatment effect, but rather a nonlinear relationship between the quantitative assignment variable and the outcome variable. When there is both a discontinuity and a change in slope between the two regression lines at the cut point, the treatment effect can be directly interpreted at the cut point (Trochim, Cappelleri, & Reichardt, 1991).

Statistical Power.

The RD design typically will have considerably lower power than the

randomized experiment, with the degree to which power is reduced depending on the extremity of the cut point and the magnitude of the correlation between the quantitative assignment variable and the posttest. Goldberger (1972) estimated that in properly specified models in which the quantitative assignment variable and the posttest had bivariate normal distributions and a relatively strong correlation, 2.75 times as many participants would be required using the RD design to achieve the same level of statistical power as in the randomized experiment. Researchers planning RD designs will need to use relatively large sample sizes, both to achieve adequate statistical power and to probe the statistical assumptions underlying the approach. Need-or merit-based programs are often delivered to large samples so that this desideratum can often be achieved.

Causal Generalization.

Generalization of results using the RD design is limited because causal inference is restricted to values on the assignment variable that are close to the cut point. Often this will be sufficient: The RD design is typically conducted with the populations and in the settings of interest. The cut points used are also often those for which there are supporting data (e.g., cutoffs for clinical levels of high blood pressure) or strong historical tradition for their use (e.g., Dean's List = 3.50 GPA). Thus, this limitation is often an issue more in theory than in practice because of the restricted range of generalization sought.

Summary .

The RD design provides an excellent approach when participants are assigned to conditions on the basis of a quantitative measure. From the standpoint of the Rubin causal model, potential outcomes can be envisioned in the treatment and control conditions for values on the assignment variable near the cut point. From Campbell's perspective, the RD design approximates a randomized experiment within a narrow range, it rules out most threats to internal validity, and any alternative explanation must match the obtained pattern of results at the cut point. From both the Rubin and Campbell traditions, the RD design is viewed as a strong alternative to the randomized experiment. The RD design typically can be conducted using the Units (participants), Treatments, Outcomes, and Settings to which the findings are to be generalized. On the negative side, the RD design is considerably lower in power than the randomized experiment, although this limitation is often moot because it is carried out with large samples. Two key

concerns regarding the design are (1) whether the assignment rule has been rigorously followed and (2) whether functional form has been properly estimated. Both issues can lead to bias in the estimate of causal effects; good methods have been developed for addressing each of these issues (Imbens & Lemieux, 2008). More complex variants of the RD design can be implemented: multiple baseline measures may be used to assign participants to treatment, multiple treatments may be delivered with multiple cut points, other baseline measures may be used as additional covariates, and multiple outcome variables may be collected (Shadish et al., 2002; Trochim, 1984). The RD design can be implemented with multiple potential assignment variables (Wong, Steiner, & Cook, 2012). The RD design can also be supplemented with a randomized tie breaking experiment around the cut point to provide even stronger inferences (Rubin, 1977; Shadish et al., 2002; see Aiken et al., 1998 for an empirical illustration of combining these designs).

Interrupted Time Series Design

The interrupted time series (ITS) design is a second quasi-experimental design that can potentially permit strong causal inferences. Once again, a quantitative assignment variable is used to assign units to treatment and control conditions, but here the assignment variable is time. In the basic ITS design, measurements of the outcome variable are collected at equally spaced intervals (e.g., daily, monthly, yearly) yielding a large number of observations over time. An intervention is implemented at a specific point in time; the intervention is expected to affect the repeatedly measured outcome variable series. The intervention may be applied (e.g., change from a fault-based to a no-fault divorce law in a state; Mazur-Hart & Berman, 1977) or it may be of more basic interest (e.g., changing the response format on a weekly survey; changing the method of participant recruitment). Time series analysis permits the researcher to identify changes in the level and slope of the series that occur as a result of the intervention.⁶ If (a) time is assumed to be a good proxy for the actual rule upon which treatment and control conditions are assigned (Judd & Kenny, 1981) and (b) the correct functional form of the relationship between time and the outcome variable is specified, then controlling for time will lead to an unbiased estimate of the treatment effect. If time is not an adequate proxy, then estimates of the magnitude of the treatment effect might be biased. Then other features of the situation or the design that covary with time might also change at the intervention point, raising potential threats to internal validity.

ITS designs have most commonly been utilized in two different research contexts: (1) societal interventions and (2) single-subject designs (Franklin, Allison, & Gorman, 1996; Shadish, Hedges, Pustejovsky, Rindskopf, Boyajian, & Sullivan, in press). As an illustration of (1), Hennigan *et al.* (1982) tested the hypothesis that the introduction of television broadcasting in U.S. cities in the late 1940s and early 1950s would lead to subsequent increases in certain crime rates. They found an increase in burglary rates, possibly associated with the common depiction of upper-middle-class lifestyles in early television programs. As an illustration of (2), Hoepfner, Goodwin, Velicer, Mooney, and Hatsukami (2008) used time series methods to understand individual differences in smokers (15–45 cigarettes per day at baseline) in their long-term responses to a smoking reduction program. An initial 14-day baseline observation period was followed by a staged 42-day nicotine replacement therapy program, which, in turn, was followed by a 40-day follow-up observation period. The program produced a substantial reduction in smoking among the participants to an average of approximately eight cigarettes per day at its end. Of particular interest, there were strong individual differences in smoking during the post-program follow-up period. Forty-seven percent of the participants showed a pattern of increased smoking over time, 40% continued to show further decreases in smoking over time, and 12% remained at their new lower post-program levels. In both (1) and (2), the treatment was assigned on the basis of time, and possible treatment effects were inferred from a change in the level of behavior that occurred between the baseline and treatment periods. Figure 4.4 illustrates some of the different types of effects that can be detected using the ITS design.

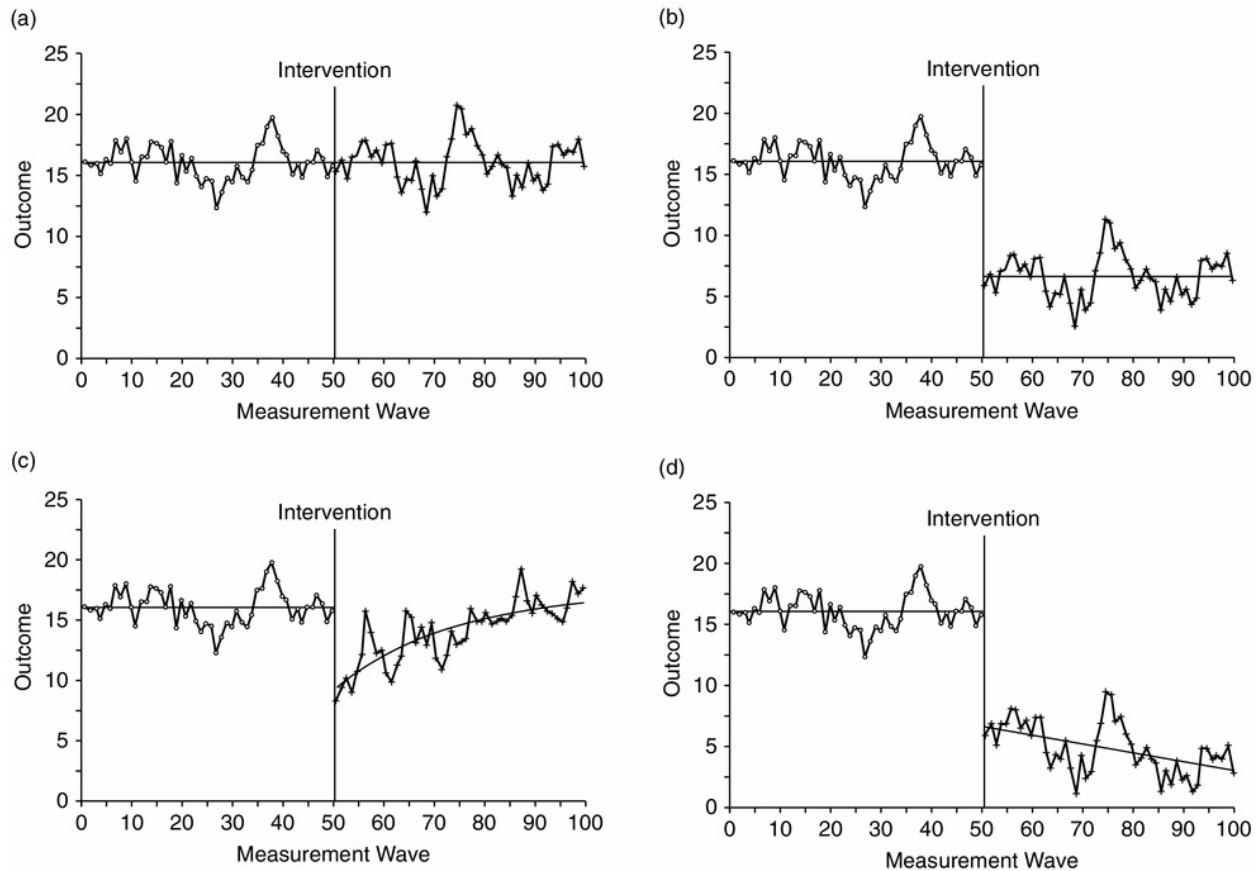


Figure 4.4. Illustration of possible outcomes in the interrupted time series design: (A) No effect of intervention. (B) Intervention decreases the level of the outcome variable. (C) Intervention produces an immediate drop in the level, followed by a slow return to the baseline level of the series. (D) Intervention produces both an immediate drop in the level and a change in the slope of the series.

Note: The same outcome is assessed at each wave of measurement of the study. The vertical line indicates the point at which intervention is introduced (between measurements 50 and 51).

From consideration of [Figures 4.3](#) and 4.4, there is a key similarity between the basic RD and ITS designs. Both designs use a quantitative assignment rule in which treatment is assigned beginning at a specific point represented on the X-axis. From Campbell's perspective, any alternative explanations of the results need to provide an account of why the level, slope, or both of the series change at that specific point – a very specific pattern of results must be matched.

Identifying Threats to Internal Validity.

Associated with the use of time as the quantitative assignment variable, three threats to internal validity must be given careful consideration given the potential association of time with other processes. Consider the example of an adolescent smoking prevention program implemented at the beginning of a school year. Based on school records, the total number of students who are cited for smoking by school personnel each month on school grounds for five years prior to and five years after the beginning of the intervention constitute the outcome series.

1. *History.* Some other event that could also be expected to decrease smoking may occur at about the same time that the smoking prevention program is implemented. For example, the community may concurrently institute fines on businesses that sell cigarettes to minors.
2. *Selection.* The population of the students in the school may change in the direction of having a greater proportion of nonsmokers. Children who smoke may transfer to other districts where the prevention program is not being implemented.
3. *Instrumentation.* Aspects of the record-keeping procedures may change, leading to decreased reports of smoking. Budget cuts may decrease the staff available for monitoring the school's grounds, leading to fewer citations.

Once again, these threats are plausible only if they occur at about the point of the intervention. For example, if the community were to introduce fines for selling cigarettes to minors three years after the implementation of the smoking prevention program, this action would *not* provide an alternative explanation for any drop in the number of students who are cited for smoking at the point of implementation of the program. These three threats to internal validity are more plausible in ITS designs evaluating societal interventions than in single-subject designs in which the experimenter has more control over the procedures and environment (Kazdin, [2011](#)).

Design Elements and Pattern Matching.

The three potential threats to internal validity can be made less plausible through the addition of one or more design elements to the basic time series design. To illustrate three design elements that have been used, we consider a study of the effect of a ban on indoor smoking on coronary heart disease admissions in Ohio

hospitals (Khuder et al., 2007). In their basic ITS design, Khuder and colleagues examined monthly hospital admissions over a 6-year period in one city that banned indoor smoking following year 3 of the study. Following the introduction of the ban, there was a drop in the number of hospital admissions for coronary heart disease in the city.

1. *No Treatment Control Series.* Comparable data may be available from another similar unit that did not receive treatment during the same time period. Khuder *et al.* identified a second comparable city in Ohio that did not institute an indoor smoking ban during the 6-year study period. No drop in hospital admissions for coronary artery disease occurred in the comparison city following year 3, the point when the ban was introduced in the treatment city.
2. *Other Control Series.* Sometimes data are available from another series that would *not* be expected to be influenced by the treatment, but which would be expected to be impacted by many of the same threats to internal validity. Khuder *et al.* examined hospital admissions over the same 6-year period for medical diagnoses that were expected to be unrelated to smoking. In both the treatment city that introduced the smoking ban and the no-treatment comparison city there was no decrease in hospital admissions for these diagnoses after year 3.
3. *Switching Replications.* In some cases, the same intervention may be implemented in more than one locale but at different times. Imagine that in the Khuder *et al.* study a third comparable city could be located that introduced the indoor smoking ban following year 4.5 of the study. If a comparable drop in hospital admissions for coronary artery disease occurred in city 3 at the point of its intervention, this would provide further evidence for the effectiveness of the indoor smoking ban. West, Hepworth, McCall, and Reich (1989) studied the effects of the implementation of a law mandating a 24-hour jail term for driving under the influence of alcohol, which was implemented at different times in Phoenix, Arizona and in San Diego, California. Analysis of highway fatality data showed an immediate 50% reduction in highway fatalities at the point of the intervention followed by a slow decrease in the magnitude of the effect over time. The replication of the same effect at different time points helps rule out other possible alternative explanations (e.g., history).

Multiple design elements can and should, when possible, be combined in a

single study to strengthen further the causal inferences that may be made. Khuder *et al.* (2007) used a no-treatment control series and another control series. West *et al.* (1989) combined all three of the design elements. In single subject designs, when the researcher has control over treatment delivery, the design can be further strengthened by introducing and removing treatments following an a priori or randomized schedule (Kazdin, 2011; Kratochwill & Levin, 2010). The specificity of the pattern of results that must be accounted for in ITS studies with multiple design elements leads to strong causal inferences, sometimes with a certainty approaching that of a randomized experiment.

Delayed Effects.

ITS designs introduce a complication relative to RD designs because treatment effects can be delayed in time, typically for one of two reasons. (a) New programs often require time for personnel to be trained before they fully go into effect; new innovations often require time to diffuse through society or for the change in policy to become known. In their study of the effects of the introduction of television, Hennigan *et al.* (1982) collected supplementary data on the proportion of households with television sets each month following the introduction of television broadcasts in the study cities to strengthen their potential causal inferences. (b) Some interventions affect processes whose outcomes will only be evident months or years later. A birth control intervention would be expected to affect birth rates beginning approximately nine months later. Theory about the process (here, human gestation periods) provides a strong basis for expecting the effect over a specified time lag distribution following the intervention. In the absence of strong theory or data on the implementation process, similar strong causal inferences cannot be made about delayed effects of interventions.

Statistical Issues in Modeling Time Series.

The basic statistical model for analyzing the ITS design parallels that of the basic RD design: $Y = b_0 + b_1\text{Time} + b_2\text{Treatment} + e$. Three complications arise from the use of Time in the analysis. First, any linear or nonlinear trends must be removed from the outcome series. Second, data over time often involve weekly, monthly, or yearly cycles that must be removed. A study of consumption of alcoholic beverages in college students must take account of the typical increases in student drinking that regularly occur on weekends. Third, adjacent observations are often more similar than observations that are further removed in

time (consider predicting tomorrow's weather from today's weather versus the weather seven days ago). This problem, known as serial dependency, implies that the residuals (e 's) in the regression equation will *not* be independent (Judd & Kenny, 1981; West & Hepworth, 1991), violating an important assumption of statistical tests. These problems can be eliminated through appropriate modeling procedures that typically involve adding terms to the regression equation or transforming the regression equation to address the problem. Judd and Kenny (1981) and West and Hepworth (1991) present introductions to the analysis of time series data; Box, Jenkins, and Reinsel (2008), Chatfield (2004), and Velicer and Molenaar (2013) offer more advanced presentations.

Statistical Power.

Two issues of statistical power commonly arise in time series analysis. First, researchers often evaluate the effects of an intervention shortly after an intervention has been implemented, giving rise to a series with few post-intervention observations. The statistical power of the analysis to detect treatment effects is greatest when the intervention occurs at the middle of the series. Second, serial dependency has complex effects on statistical power. To the degree that observations are not independent, each observation in effect counts less than 1, increasing the total number of observations needed to achieve a specified level of statistical power. On the other hand, the use of the same participant or same population of participants throughout the series reduces the variability of the series relative to one in which different participants are sampled at each time point. The statistical power of a given time series analysis is determined in large part by the trade-off between these two effects in the specific research context.

Summary and Conclusion.

ITS designs provide a strong quasi-experimental approach when interventions are introduced at a specific point in time. Such designs provide a strong basis for the evaluation of the effects of changes in social policy or interventions that are implemented at a specific point in time. The basic time series design often allows researchers to credibly rule out several threats to internal validity: Any alternative explanation, to be plausible, must account for why the change in the series occurred at a particular point in time. The use of design elements such as a comparable no-treatment site, control series that are expected to be affected by the threats to interval validity but not the treatment, switching replications, and

the introduction and removal of treatments can further strengthen causal inference. Given their ability to rule out alternative explanations through both design elements and statistical adjustment, ITS designs represent one of the strongest alternatives to the randomized experiment.

Nonequivalent Control Group Designs (Observational Studies)

The most common alternative to the randomized experiment is the nonequivalent control group design (a.k.a, observational study; Cochran, 1965). In this design, a group of participants is given a treatment or a “natural event” (e.g., earthquake) occurs to the group. A second (comparison) group that does not receive the treatment is also identified. The treatment and comparison group participants are measured both at baseline and following treatment. The “assignment” process through which participants end up in the treatment and comparison groups is unknown and presumed to be nonrandom. The goal is to infer the causal effect of the treatment. For example, Lehman, Lampert, and Nisbett (1988) compared the level of statistical reasoning of advanced graduate students in psychology (which emphasizes training in statistics) with that of advanced graduate students in other disciplines (e.g., chemistry) that have a substantially lower training emphasis in statistics. The outcomes in the treatment groups at posttest were compared after attempting to remove any differences between the groups that were measured at pretest.

From the perspective of Rubin Causal Model, the central challenge of the nonequivalent control group design is to assure that the treatment and control groups are equated at baseline, given the absence of randomization. Rosenbaum (2010) identifies two classes of baseline variables. Covariates are variables assessed at baseline and hidden variables are *not* assessed at baseline. If (1) the treatment and control groups can be equated on all baseline variables that are potentially related to both treatment condition and outcome and (2) each unit (participant) has some probability of being in either the treatment or control groups (i.e., $0 < \text{probability of treatment} < 1$) – a condition known as *strong ignorability* – then an unbiased estimate of the causal effect is possible. However, given that some of the baseline differences may be on hidden variables, researchers may legitimately retain considerable uncertainty as to whether all preexisting differences between the treatment and control groups on covariates and hidden variables have been adequately ruled out as potential explanations for observed differences on the outcome variable. Major disputes

have occurred in the literature over the true effectiveness of treatments that have been evaluated using observational studies, for example the Head Start Program (e.g., Barnow, 1973; Bentler & Woodward, 1978; Cicarelli, Cooper, & Granger, 1969; Magidson, 1977 for diverse analyses and re-analyses of the Head Start Program data).

Strategies for Equating Groups.

From the standpoint of the Rubin Causal Model, two strategies for equating the groups may be undertaken. (a) Mimicking the RD design, the researcher can model the mechanism by which participants were assigned to the treatment and control groups. When this goal is successfully accomplished, the treatment effect estimate will be an unbiased estimate of the causal effect (Cook, Shadish, & Wong, 2008; Morgan & Winship, 2007). (b) In the absence of information about the assignment mechanism, the researcher can attempt to model selection into treatment and control groups using propensity scores (defined later in the chapter). In this approach, the researchers measure a rich set of baseline variables chosen based on theory and prior research. The key baseline variables that should be measured are those that may be related to *both* (a) selection into the treatment and control groups and (b) the outcome variable in the population. Such variables can produce a spurious relationship between the treatment and the outcome by serving as a common cause of both. For example, at the beach the baseline variable of outdoor temperature serves to produce a correlation between consumption of ice cream and swimming, even though the “treatment” of ice cream consumption (yes vs. no) does not lead to the “outcome” of increased swimming. In contrast, baseline variables that are only related to either (a) or (b), but not both, do *not* bias treatment effect estimates (Berk, 1988; Pearl, 2009). If there were either no link between temperature and ice cream consumption or no link between temperature and swimming, the spurious relationship between ice cream consumption and swimming would not occur. A good proxy for key baseline variables in some, but not all, research contexts is a pretest variable that is operationally identical to the outcome variable, particularly if the period between the pretest and posttest measurements is not long (Reichardt & Mark, 2004).

As in the RD design, the estimate of the treatment effect needs to be conditioned on the set of baseline covariates. Rosenbaum and Rubin (1983) developed the propensity score as a method of summarizing all of the information contained in a large set of covariates as a single number. The

propensity score is the predicted probability that the participant will be assigned to the treatment group based on the set of measured covariates. Typically propensity scores are estimated using logistic regression where the outcome is treatment assignment and the full set of covariates is used as the predictors, possibly also including curvilinear effects or interactions of the predictors. If (a) propensity scores can be successfully constructed and (b) treatment and control groups can be successfully equated on the propensity scores, then theoretically the groups will also be equated on all variables that enter in to the propensity score. Several methods including matching, blocking, weighting, and analysis of covariance can be used to equate treatment and control groups on propensity scores. We focus on matching below, which can have advantages over other approaches in certain contexts (Schafer & Kang, 2008). We first illustrate how matching works in its simplest form using a single covariate and then introduce propensity scores that provide a method of matching on many variables simultaneously. West, Cham, Thoemmes, Renneberg, Schultze, and Weiler (in press) provide an introduction to the use of propensity scores.

Consider an observational study comparing two school classrooms ($n_A = 12$; $n_B = 13$), one of which is to be given a new instructional treatment and one of which is to be given standard instruction (control group). Our matching variable IQ is the only variable believed to be related to both selection into the treatment and control classrooms and educational outcome. Table 4.4 presents hypothetical data for this illustration in which scores on the covariate IQ have been ordered from low to high within each group. Adequate matches are available for 10 pairs of students. Following matching, the mean difference on the baseline IQ for the 10 matched pairs is considerably smaller ($\bar{X}_A - \bar{X}_B = 0.1$) than for the full, unmatched classes ($\bar{X}_A - \bar{X}_B = 11$). Note that adequate matches are not available for all participants. The children with the two highest and three lowest IQ scores must be dropped from the analysis; consequently, generalization outside the range of successful matching (here, IQ = 97 to 128) will be limited.

Table 4.4. Illustration of Simple Matching in Two Classrooms on Pretest IQ Scores

Pair	Classroom A	Classroom B
	150	
	130	

	130	
1	125	128
2	120	119
3	118	119
4	117	116
5	115	116
6	110	112
7	108	106
8	103	102
9	100	99
10	99	97
		92
		85
		80

Note: Scores are ordered within classes and represent the pretest IQ measures of the students. Pairs of students on the same line represent matched pairs. Ten pairs of students were matched. Two students in Classroom A and three students in Classroom B have no matched pair and are thrown out of the design. The mean IQ for all 12 students in Classroom A is 116; the mean IQ for all 13 students in Classroom B is 105. For the 10 matched pairs, the mean IQ is 111.5 in Classroom A and 111.4 in Classroom B.

Now consider attempting to match treatment and comparison groups on 10, 20, or more such covariates. The task becomes overwhelming because the matching involves too many variables. Instead, we use propensity scores because they can reduce all of the information in the set of measured baseline covariates

related to treatment assignment to a single summary number – the propensity score. The researcher can then match on the basis of the single propensity score instead of the large set of covariates. If the propensity score has been properly constructed, the two groups will be approximately equated not only on the propensity score, but on each of the covariates that was used in the estimation of the propensity score. Matching is then accomplished using modern software (e.g., MatchIt; Ho, Imai, King, & Stuart, 2011) that optimally pairs participants in each group on the propensity score. Constraints can be included to assure that no match in the sample will exceed a specified small difference in propensity scores (e.g., .025), termed a caliper. In some cases the comparison group will be much larger in size than the treatment group: More than one control group member can be closely matched with each treatment group member to maximize the sample size and hence the power of statistical tests (see Ming & Rosenbaum, 2000; Stuart & Green, 2008).

Finally, the use of matched samples permits checks on the quality of the equating of the two groups. The treatment and control groups can be compared on each of the baseline variables that were used to construct the propensity score. To the extent that matching has been successful, comparison of treatment and control groups on baseline covariates should *not* show statistically significant differences and differences should be small in magnitude, paralleling a randomized experiment. Moser, West, and Hughes (2012) showed that following matching on propensity scores, matched, low-achieving children who were promoted versus retained at the end of first grade significantly differed on fewer than 5% of the 72 covariates used to construct the propensity score. None of the baseline differences exceeded a small effect size.⁷

In summary, matching on propensity scores can greatly reduce selection bias and thereby strengthen causal inference by equating the treatment and control groups at baseline. Matching offers the theoretical advantages of minimizing extrapolation by only comparing participants who have similar baseline scores and by making the proper adjustment regardless of the functional form of the relationship between the covariates and outcome (Rosenbaum, 1986). It also allows for checks on the propensity score procedure by allowing checks on the equivalence of each of the baseline covariates in the two groups. Matching's chief liability is that there is no guarantee that all key covariates have been measured at pretest; matching does *not* equate groups on hidden variables. Often overlooked are covariates related to the participant's relative preference for the control and active treatments (Imbens, 2010) and the possibility that participants have different growth rates prior to treatment (Haviland, Nagin, & Rosenbaum,

2007). The careful choice of a comprehensive selection of reliably measured covariates is the most important determinant of the success of the baseline equating of the groups (Cook & Steiner, 2010; Shadish, 2013). The causal effect estimated by matching on propensity scores can appropriately be generalized only to the population of participants that could potentially receive either the control or active treatments (the average treatment effect for the treated participants). Alternative weighting methods that use the propensity scores to weight each case permit estimation of the ACE for the full population (West et al., 2013).

Threats to Internal Validity.

In terms of Campbell's perspective, the strength of the nonequivalent control group design is that the inclusion of the control group rules out basic threats such as maturation that occur equally in the treatment and control groups. The weakness of the design is that it does not rule out interactions with selection in which these threats operate differently in the treatment and control groups.

1. *Selection × Maturation.* The treatment and control groups may differ in the rate at which the participants are growing or declining on the outcome variable prior to treatment. Suppose a smoking prevention program were given to all students in a suburban high school, whereas students in an urban high school received a control program. Selection × maturation would be a threat if the number of cigarettes smoked per day in the urban high school were increasing at a faster rate than in the suburban high school in the absence of intervention.
2. *Selection × History.* Some other event may occur to only one of the two groups that might affect the outcome variable. Suppose the media in the suburban but not the urban site started a series pointing out the dangers of teenage smoking. Or, lower-cost contraband cigarettes became available in the urban but not in the suburban area during the study.
3. *Selection × Statistical Regression.* Participants may be selected for the treatment or control group based on an unreliable or temporally unstable measured variable. Participants selected for the study in the inner-city school on average may have been temporarily smoking fewer than their normal number of cigarettes per week (e.g., if several of the participants had just recovered from colds), whereas participants in the suburban school on average may have been smoking their normal number of cigarettes per day. Upon remeasurement in the absence of

treatment, participants in each group would tend to report their typical level of smoking.

4. *Selection × Instrumentation*. Some aspect of the measuring instrument may change from pretest to posttest in only one of the two groups. This threat can take many forms, which can be manifested in such problems as differences between the two groups in the reliability of the scale, or ceiling or floor effects in the scale. Local changes in record-keeping practices or in the sensitivity of the measures can also produce this threat.

Design Elements and Pattern Matching.

The basic nonequivalent control groups design may also be strengthened by adding design elements that specifically address the plausible threats to validity. When the pattern of results matches that expected from the treatment rather than that expected if the specific threats to internal validity exist, causal inference is strengthened. Shadish *et al.* (2002, p. 157) present an extensive list, and [Figure 4.5](#) presents an observational study that illustrates their use.

Multiple Control Groups.

Typically, no control group can be identified that is comparable to the treatment group on all factors that could affect the outcome of the study. It is often possible, however, to identify multiple imperfect control groups, each of which can address some of the threats to validity. For example, Roos, Roos, and Henteleff (1978) compared the pre-and post-operation health of tonsillectomy patients with (a) a nonoperated control group matched on age and medical diagnosis and (b) nonoperated siblings who were within five years of the patient's age. The first control group roughly equates the medical history of the treatment participant; the second control group roughly equates the genetic predisposition, shared environment, and family influences of the treatment participant (Lahey & D'Orofrio, 2010). Obtaining similar estimates of the causal effects across each of the comparison groups can help minimize the plausibility of alternative explanations of the obtained results. Rosenbaum (1987; 2002) discusses several methods of using data from carefully chosen multiple control groups to reduce the likelihood that hidden variables may be responsible for the results.

Nonequivalent Dependent Variables.

Sometimes data can be collected on additional dependent variables or in different contexts on the same group of participants. Ideally, the measures selected should be conceptually related to the outcome variable of interest and should be affected by the same threats to internal validity as the primary dependent variable. However, the researcher would *not* expect these measures to be affected by the treatment. Roos *et al.* (1978) compared the tonsillectomy and two control groups on their primary outcome variable, health insurance claims for respiratory illness, which they expected to decrease only in the treatment group. They also compared health insurance claims for other nonrespiratory illness that would *not* be expected to decrease following treatment. To the extent that the hypothesized pattern of results is obtained on the primary outcome variable but not on the nonequivalent dependent variables, the interpretation of the results as a causal effect is strengthened.

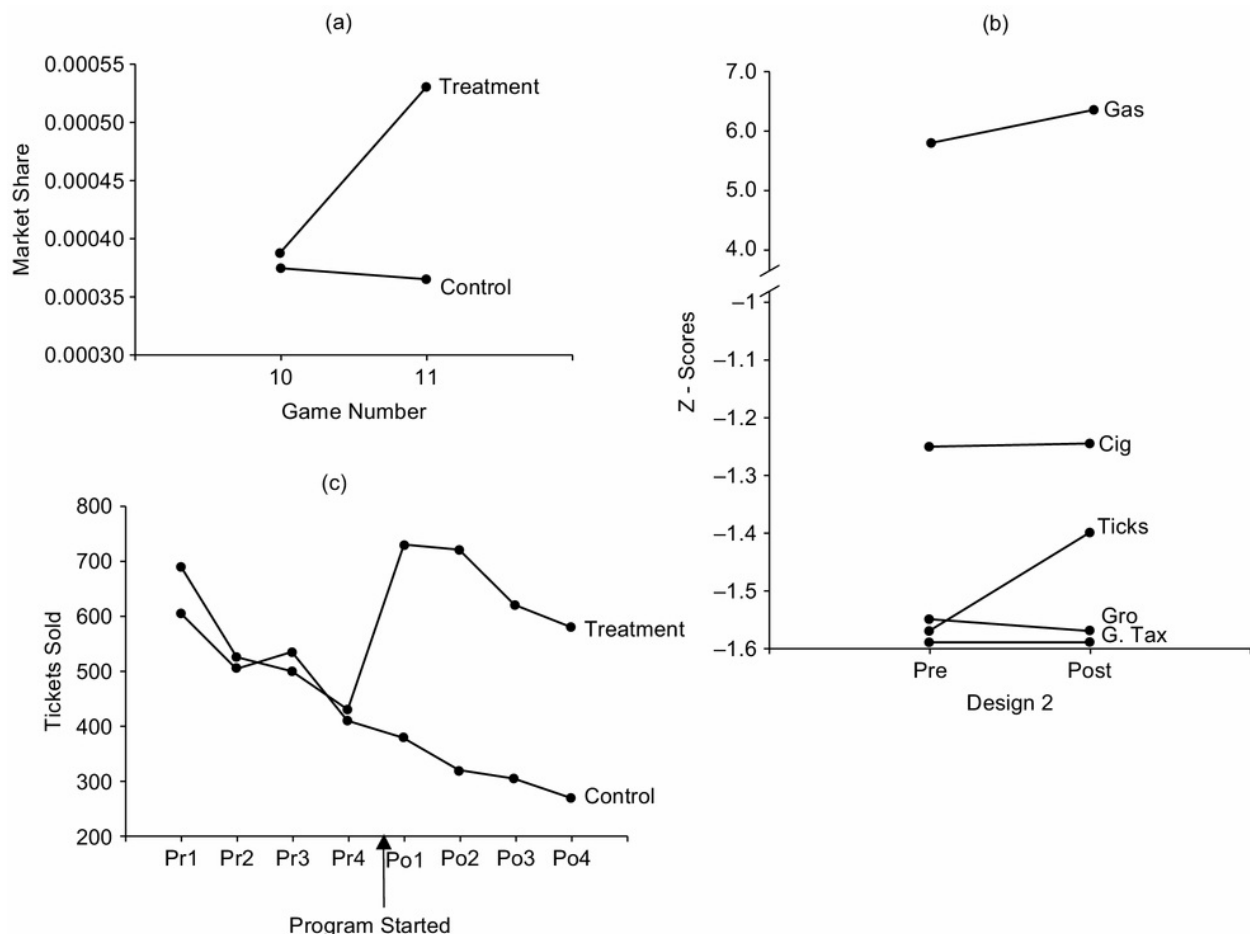


Figure 4.5. Using design elements to strengthen causal inferences in observational studies: (a) Matching. Treatment and control stores are selected from the same chain, are in the same geographical location, and are comparable in sales during baseline (lottery game 10). Introduction of the treatment during

lottery game 11 yields an increase in sales only in the treatment stores. (b) Nonequivalent Dependent Variables. Within the treatment stores, sales of lottery tickets increase substantially following the introduction of treatment. Sales of other major categories – gasoline, cigarettes, groceries (nontaxable), and groceries (taxable) – that would be expected to be affected by confounding factors, but not treatment, do not show appreciable change. (c) Repeated Pre-and Posttest Measurements. Treatment and control stores sales show comparable trends in sales during the four weeks prior to and following the introduction of the treatment. The level of sales in the treatment and control stores is similar prior to the introduction of treatment, but differ substantially beginning immediately after treatment is introduced. Adapted from Reynolds, K. D. & West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review*, 11, 691–714.

Multiple Pretests over Time.

The addition of multiple pretests on the same variable over time to the nonequivalent control group design can help rule out threats associated with selection \times maturation or selection \times regression in the two groups. In this design, the multiple pretests are used to estimate the rates of maturation in the two groups prior to treatment. These estimates can then be used to adjust the treatment effect under the assumption that the pattern of maturation within each group would not change during the study. Figure 4.5(c) illustrates how the use of multiple pretests on an operationally identical outcome variable can adjust for potential (maturation) trends in sales levels.

Statistical Power.

Whereas in the randomized experiment techniques like matching and analysis of covariance almost always increase power, these techniques are used in the observational study primarily to reduce bias – they may increase, decrease, or have no effect on statistical power. To the extent that the condition of strong ignorability can be approached, the nonequivalent control group design theoretically provides an unbiased estimate of the causal effect. In practice, it is difficult for researchers to convincingly establish that the assumptions necessary for strong ignorability have been met, so multiple analyses must often be conducted to rule out plausible threats to internal validity. Methodologists frequently suggest that researchers conduct diverse analyses with different underlying assumptions (sensitivity analyses) in an attempt to bracket the

estimate of the causal effect (Reichardt & Gollob, 1997; Rosenbaum, 2002; 2010). This strategy increases the certainty of our causal inference to the extent that all of the analyses yield similar estimates of the treatment effect. However, this strategy also implies that if we wish to show that the bracket does not include a treatment effect of 0, our estimate of statistical power must be based on the statistical model that can be expected to lead to the smallest estimate of the magnitude of the treatment effect. As a result, the power of the statistical tests in the nonequivalent control group design can be expected to be lower than for a randomized experiment.

Causal Generalization.

Observational studies are normally conducted with the units, treatments, settings, and observations of interest and are thus potentially high in causal generalization. Often they are conducted when randomization is not possible for ethical or practical reasons so that this design is the only feasible option. In other cases, randomization is not accepted by the community (e.g., randomization to a faith-based versus secular drug treatment program), so that participants and treatment settings in any randomized experiment would be highly atypical of the UTOS to which the researchers intend to generalize their results (West et al., 2008). In such cases, observational studies can be a valuable supplement to randomized experiments. The ability to generalize causal effects obtained in well-designed observational studies to UTOS of interest is one of the design's primary strengths; the inability to make precise statements about the magnitude (and sometimes direction) of causal effects is its primary weakness.

Summary and Conclusion.

In terms of causal inference, the basic nonequivalent control group design provides the least satisfactory of the three quasi-experimental alternatives to the randomized experiment we have considered. From the perspective of the Rubin Causal Model, there is always some uncertainty as to whether the condition of strong ignorability has been met. From Campbell's perspective (e.g., Shadish et al., 2002), four threats to internal validity resulting from interactions with selection must be ruled out. Internal validity of the design can potentially be enhanced by the inclusion of design elements and pattern matching. Data may sometimes be collected that permit the researcher either to model adequately the selection process or match groups on propensity scores. Strong treatment effects that can be shown through sensitivity analyses to be robust to the potential

influence of hidden variables coupled with careful consideration of the threats to internal validity in the research context have the potential to overcome the limitations of this design.

How Well Do the Alternative Nonrandomized Designs Work?

Two different approaches have been taken to comparing the results of nonrandomized quasi-experimental designs and experimental designs. First, the effect sizes from published articles that combined one or more of the three quasi-experimental designs with a randomized experiment within the same study have been compared (Cook *et al.*, 2008; Cook & Wong, 2008). Studies comparing randomized experiments and regression discontinuity designs and randomized experiments and interrupted time series designs that share the same treatment condition have found very similar effect sizes. In contrast, similar studies comparing the results randomized experiments with observational studies have found more variable results (see also West & Thoemmes, 2008). Cook *et al.* (2008) conclude that similar effect sizes will be found when the selection processes is known or measured or the treatment and control groups have been sampled to be very similar. The similarity of the results is enhanced when a rich set of baseline covariates are carefully selected and measured so that proper adjustment of the treatment effect in the observational study can be undertaken. In addition, similar control procedures other than randomization need to be used in the randomized experiment and observational study (West, 2008).

Second, Shadish, Clark, and Steiner (2008) and Pohl, Steiner, Eisermann, Soellner, and Cook (2009) randomly assigned participants to a randomized experiment or observational study with the same treatment and control conditions. In the observational study, participants chose their preferred treatment condition. Given adjustment of the treatment effect for a carefully selected set of pre-treatment covariates in the observational study, similar effect sizes were obtained in both designs. A similar experiment in which participants were randomly assigned to a randomized experiment or a regression discontinuity design involving identical treatments also produced similar estimates of effect sizes (Shadish, Galindo, Wong, Steiner, & Cook, 2011). Taken together, these studies indicate that randomized experiments and carefully conducted quasi-experiments have a strong potential for producing the same results. These findings are currently limited by the small number of published studies making within study comparisons between designs and the limited

contexts and treatments to which the randomized comparisons have been applied.

Some Final Observations

This chapter has provided an introduction to experimental and quasi-experimental designs that are useful in field settings. In contrast to the laboratory research methods in which social psychologists have traditionally been trained, modern field research methodology reflects the more complex and less certain real-world settings in which it has been applied (Cialdini & Paluck, [Chapter 5](#) in this volume). Well-designed randomized experiments often turn into “broken randomized experiments” because of issues like attrition and treatment noncompliance. Quasi-experimental designs, such as the basic observational study, rule out a more limited set of the threats to internal validity. Researchers working in field settings must carefully identify potential threats to internal validity. They need to add design elements such as multiple comparison groups and multiple pretest tests that address the identified threats and violations of assumptions. Appropriate statistical models may also be needed to address specific threats. The use of these procedures can in some cases produce strong designs for which the certainty of causal inference approaches that of a randomized experiment. In other cases, threats to internal validity may remain plausible despite the investigator's best efforts. [Table 4.5](#) presents a summary of the violations of assumptions, threats to internal validity, and design and statistical modeling approaches that address these issues.

Table 4.5 *Key Assumptions/Threats to Internal Validity: Example Remedies for Randomized Experiments and Quasi-Experiments*

Assumption or Threat to Internal Validity	Approaches to Mitigating the Threat	
	Design Approach	Statistical Approach
Randomized Controlled Experiment		
Independent units	Temporal or geographical isolation of units	Multilevel analysis; Other statistical

		adjustment for clustering
Stable Unit Treatment Value Assumption (SUTVA): Other treatment conditions do not affect participant's outcome; No hidden variations in treatments	Temporal or geographical isolation of treatment groups	Statistical adjustment for measured exposure to other treatments
Full treatment adherence	Incentives for adherence	Local average treatment effect
No attrition	Sample retention procedures	Missing data analysis
Regression Discontinuity Design		
Functional form of relationship between assignment variable and outcome is properly modeled	Replication with different cutpoint; Nonequivalent dependent variable	Nonparametric regression; Sensitivity analysis
Interrupted Time Series Analysis		
Functional form of the relationship for the time series is properly modeled;	Nonequivalent control series in which intervention is not introduced; Switching	Diagnostic plots (lowess; autocorrelogram)
Another historical event, a change in population (selection), or a change in measures coincides with the introduction of the	replication in which intervention is introduced at another time point; Nonequivalent	Sensitivity analysis

the introduction of the intervention.

Nonequivalent dependent measure.

Observational Study

Measured baseline variables equated

Multiple control groups

Propensity score analysis

Unmeasured baseline variables equated

Nonequivalent dependent measures

Sensitivity analysis

Differential maturation

Additional pre-and post-intervention measurements

Subgroup analysis

Note: The list of assumptions/threats to internal validity identifies issues that commonly occur in each of the designs. The alternative designs may be subject to each of the issues listed for the randomized experiment in addition to the issues listed for the specific design. The examples of statistical and design approaches for mitigating the threats to internal validity illustrate some commonly used approaches and are not exhaustive. For the observational study design, Rubin's and Campbell's perspectives differ so that the statistical and design approaches do not map 1:1 onto the listed assumptions/threats to internal validity. Adapted from West, S. G. (2009), *Current Directions in Psychological Science*, 18, 199–230.

Because of the complexity and uncertainty associated with research conducted in the field, it is important for researchers to acknowledge publicly the known limitations of their findings and to make their data available for analysis by other researchers (Funder, Levine, Mackie, Morf, Vazire, & West, in press; Houts et al., 1986; Sechrest et al., 1979). In the Campbell tradition, public criticism of research provides an important mechanism through which biases and threats to internal validity can be identified and their likely effects, if any, on the results of a study can be assessed (West & Thoemmes, 2010). Additional studies, with different strengths and weaknesses, can then be conducted to help bracket the true causal effect. Although considerable uncertainty may be associated with the results of any single study, meta-analytic studies may reveal a consistent finding in body of studies having different strengths and weaknesses that can potentially increase confidence in the robustness of the causal effect (Johnson & Eagly,

Chapter 26 in this volume; Matt & Cook, 2009; Shadish et al., 2002).

Reflecting practice, the majority of the research examples discussed in this chapter have been applied in nature. The development of the methods discussed in this chapter has provided a strong basis for applied social psychologists to make causal inferences about the effects of treatment programs delivered in settings of interest in the field (Shadish & Cook, 2009). Some basic research in social psychology and personality now focuses on areas such as the influence of culture, major life stressors, religion, intimate relationships, and evolution of social behavior. Some questions in these areas may not be well answered by the traditional modal study – a randomized experiment, conducted in the laboratory, lasting no more than an hour, and using undergraduate students as participants (West et al., 1992). These questions call for new variants of the laboratory experiment that incorporate some of the features of modern field experiments and quasi-experiments discussed in this chapter. Some of these questions may also benefit from a possible return to improved versions of research methods more commonly used in previous eras in the history of social psychology – experiments, quasi-experiments, and other studies testing basic social psychological principles in field contexts (Bickman & Henchy, 1972; Reis & Gosling, 2010; West & Graziano, 2012). Such potential developments would help social psychology broaden the Units, Treatments, Observations, and Settings represented in its research base, providing a stronger basis for causal generalization. They would supplement the demonstrated strengths and complement the weaknesses of traditional laboratory experiments. Such designs hold the promise of a more balanced mix of methodological approaches to basic issues in social psychology in which researchers could make legitimate claims for both the internal validity and the causal generalization of their effects.

References

- Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review*, 22, 207–244.
- Aiken, L. S., West, S. G., Woodward, C. K., & Reno, R. R. (1994). Health beliefs and compliance with mammography screening recommendations in asymptomatic women. *Health Psychology*, 13, 122–129.
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal

- effects using instrumental variables (with commentary). *Journal of the American Statistical Association*, 91, 444–472.
- Baker, S. G. (1998). Analysis of survival data from a randomized trial with all-or-none compliance: Estimating the cost-effectiveness of a cancer screening program. *Journal of the American Statistical Association*, 93, 929–934.
- Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics*, 6, 267–285.
- Barnow, L. S. (1973). The effects of Head Start and socioeconomic status on cognitive development of disadvantaged students (Doctoral dissertation, University of Wisconsin). *Dissertation Abstracts International*, 1974, 34, 6191A.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Bentler, P. M., & Woodward, J. A. (1978). A Head Start re-evaluation: Positive effects are not yet demonstrable. *Evaluation Quarterly*, 2, 493–510.
- Berk, R. A. (1988). Causal inference for sociological data. In N. J. Smeltzer (Ed.), *Handbook of sociology* (pp. 155–172). Newbury Park, CA: Sage.
- Berkowitz, L. (1993). *Aggression: Its, causes, consequences, and control*. New York: McGraw-Hill.
- Bickman, L. & Hency, T. (1972) (Eds.). *Beyond the laboratory: Field research in social psychology*. New York: McGraw-Hill.
- Biglan, A., Hood, D, Borzovsky, P., Ochs, L., Ary, D., & Black, C. (1991). Subject attrition in prevention research. In C. G. Luekfeld & W. Bukowski (Eds.), *Drug abuse prevention intervention research: Methodological issues* (pp. 213–234). Washington, DC: NIDA Research Monograph #107.
- Boruch, R. F. (1997). *Randomized experiments for planning and evaluation*. Thousand Oaks, CA: Sage.
- Boruch, R. F., McSweeney, A. J., & Soderstrom, E. J. (1978). Randomized field experiments for program planning, development, and evaluation. *Evaluation Quarterly*, 2, 655–695.
- Box, G. E. P., & Draper, N. R. (1987). *Empirical model building and response*

surfaces. New York: Wiley.

- Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis: Forecasting and control* (4th Ed.). Hoboken, NJ: Wiley.
- Braucht, G. N., & Reichardt, C. S. (1993). A computerized approach to trickle-process, random assignment. *Evaluation Review*, 17, 79–90.
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297–312.
- Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (Vol. 31, pp. 67–78). San Francisco: Jossey-Bass.
- Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Chatfield, C. (2004). *The analysis of time series: An introduction* (6th Ed.). Boca Raton, FL: Chapman & Hall.
- Cicarelli, V. G., Cooper, W. H., & Granger, R. L. (1969). *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development*. Athens: Ohio University and Westinghouse Learning Corporation.
- Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.
- Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A*, 128, 134–155.
- Cochran, W. G. & Cox, G. M. (1957). *Experimental designs* (6th Ed.). New York: Wiley.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Erlbaum.
- Collins, L. M., Murphy, S. A., Nair, V. N., & Strecher, V. J. (2005). A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine*, 30, 65–73.
- Conner, R. F. (1977). Selecting a control group: An analysis of the randomization process in twelve social reform programs. *Evaluation*

Quarterly, 1, 195–244.

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. *New Directions for Program Evaluation*, 37, 39–81.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin.

Cook, T. D., Shadish, W. J., & Wong, V. C. (2008). Three conditions under which observational studies produce the same results as experiments. *Journal of Policy Analysis and Management*, 27, 724–750.

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of covariate choice, unreliable measurement and mode of data analysis. *Psychological Methods*, 15, 56–68.

Cook, T. D., & Wong, V. C. (2008). Empirical tests of the validity of the regression discontinuity design. *Annals of Economics and Statistics*, 91/92, 127–150.

Cronbach, L. J. (1982). *Designing evaluations of social and educational programs*. San Francisco: Jossey-Bass.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London: Arnold.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, 20, 115–147.

Efron, B., & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials (with discussion). *Journal of the American Statistical Association*, 86, 9–26.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, 16, 1–16.

Evans, R. I., Rozelle, R. M., Maxwell, S. E., Raines, B. E., Dill, C. A., Guthrie, T. J., Henderson, A. H., & Hill, P. C. (1981). Social modeling films to deter smoking in adolescents: Results of a three-year field investigation. *Journal of Applied Psychology*, 66, 399–414.

- Faul, F., Erdfelder, E., Lang, A-G, & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.
- Fisher, R. A. (1935). *The design of experiments*. London: Oliver & Boyd.
- Flay, B. R. (1986). Psychosocial approaches to smoking prevention: A review of findings. *Health Psychology*, 4, 449–488.
- Frangakis, C. E., & Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, 86, 365–379.
- Fraley, R. C., & Marks, M. J. (2007). The null hypothesis significance testing debate and its implications for personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 149–169). New York: Guilford.
- Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.) (1996). *Design and analysis of single case research*. Mahwah, NJ: Erlbaum.
- Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Vazire, S., & West, S. G. (in press). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Bulletin*.
- Goldberger, A. S. (1972, April). *Selection bias in evaluating treatment effects: Some formal illustrations* (Discussion paper 123–72). Madison: University of Wisconsin, Institute for Research on Poverty.
- Gonzales, N. A., Dumka, L. E., Millsap, R. E., Gottschall, A., McClain, D. B., Wong, J. J., Mauricio, A. M., Wheeler, L., Germán, M., & Carpentier, F. D. (2012). Randomized trial of a broad preventive intervention for Mexican American adolescents. *Journal of Consulting and Clinical Psychology*, 80, 1–16.
- Graham, J. W. (2012). *Missing data: Analysis and design*. New York: Springer.
- Green, P. J., & Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach*. Boca Raton, FL: Chapman & Hall.
- Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation

- of treatment effects with a regression-discontinuity design. *Econometrica*, 69, 201–209.
- Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods*, 12, 247–267.
- Hennigan, K. M., del Rosario, M. L., Heath, L., Cook, T. D., Wharton, J. L., & Calder, B. J. (1982). Impact of the introduction of television on crime in the United States: Empirical findings and theoretical implications. *Journal of Personality and Social Psychology*, 42, 461–477.
- Henry, P. J. (2008). College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry*, 19, 49–71.
- Hirano, K., Imbens, G. W., Rubin, D. B., & Zhou, X. H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1, 69–88.
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42, 1–28.
- Hoepfner, B. B., Goodwin, M. S., Velicer, W. F., Mooney, M. E., & Hatsukami, D. K. (2008). Detecting longitudinal patterns of daily smoking following drastic cigarette reduction. *Addictive Behaviors*, 33, 623–639.
- Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, 81, 945–970.
- Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models (with discussion). In C. Clogg (Ed.), *Sociological methodology 1988* (pp. 449–493). Washington, DC: American Sociological Association.
- Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy. *Journal of the American Statistical Association*, 101, 901–910.
- Houts, A. C., Cook, T. D., & Shadish, W. R. (1986). The person-situation debate: A critical multiplist perspective. *Journal of Personality*, 54, 52–105.
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd Ed.).

New York: Routledge.

- Hunter, J. E. (1996, August). Needed: A ban on the significance test. In P. E. Shrout (chair), *Symposium: Significance tests-should they be banned from APA journals?* American Psychological Association, Toronto, Canada.
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15, 309–334.
- Imbens, G. W. (2010). An economist's perspective on Shadish (2010) and West and Thoemmes (2010). *Psychological Methods*, 15, 47–55.
- Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142, 615–635.
- Imbens, G. W., & Rubin, D. B. (in press). *Causal inference: Statistical methods for estimating causal effects in biomedical, social, and behavioral sciences*. New York: Cambridge University Press [in preparation, Department of Economics, Harvard University, Cambridge, MA].
- Jo, B. (2002). Statistical power in randomized intervention studies with noncompliance. *Psychological Methods*, 7, 178–193.
- Jo, B., Ginexi, E. M., & Ialongo, N. S. (2010). Handling missing data in randomized experiments with noncompliance. *Prevention Science*, 11, 384–396.
- Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions*. New York: Cambridge University Press.
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103, 54–69.
- Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd Ed.). New York: Oxford University Press.
- Kenny, D. A., & Judd, C. M. (1986). The consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, 82, 345–362.
- Khuder, S. A., Milz, S., Jordan, T., Price, J., Silvestri, K., & Butler, P. (2007). The impact of a smoking ban on hospital admissions for coronary heart

- disease. *Preventive Medicine*, 45, 3–8.
- King, G., Nielsen, R., Coberley, C., Pope, J. E., & Wells, A. (2011). Avoiding randomization failure in program evaluation, with application to the medicare health support program. *Population Health Management*, 14, S11–S22.
- Kish, L. (1987). *Statistical designs for research*. New York: Wiley.
- Kopans, D. B. (1994). Screening for breast cancer and mortality reduction among women 40–49 years of age. *Cancer*, 74, 311–322.
- Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, 15, 122–144.
- Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.
- Lahey, B. B., & D’Orofrio, B. M. (2010). All in the family: Comparing siblings to test causal hypotheses regarding environmental influences on behavior. *Current Directions in Psychological Science*, 19, 319–323.
- Larsen, R. J. (1989). A process approach to personality psychology: Utilizing time as a facet of data. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 177–193). New York: Springer-Verlag.
- Lee, Y., Ellenberg, J., Hirtz, D., & Nelson, K. (1991). Analysis of clinical trials by treatment actually received: Is it really an option? *Statistics in Medicine*, 10, 1595–1605.
- Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of group training on reasoning: Formal discipline and thinking about everyday events. *American Psychologist*, 43, 531–442.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd Ed.). Hoboken, NJ: Wiley.
- Lohr, S. (2010). *Sampling: Design and analysis* (2nd Ed.). Boston: Brooks/Cole.
- Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics*, 122, 159–208.

- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Lawrence Erlbaum.
- Magidson, J. (1977). Toward a causal modeling approach to adjusting for preexisting differences in the nonequivalent group situation: A general alternative to ANCOVA. *Evaluation Quarterly*, 1, 399–420.
- Mark, M. M., & Mellor, S. (1991). Effect of self-relevance of an event on hindsight bias: The foreseeability of a layoff. *Journal of Applied Psychology*, 76, 569–577.
- Matt, G. E. (2003). Will it work in Münster? Meta-analysis and the empirical generalization of causal relationships. In R. Schulze, H. Holling, & V. Böhning (Eds.), *Meta-analysis: New developments and applications in medical and social sciences* (pp. 113–128). Cambridge, MA: Hogrefe & Huber.
- Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *Handbook of research synthesis and meta-analysis* (2nd Ed., pp. 537–560). New York: Russell Sage.
- Mayer, A., Thoemmes, F., Rose, N., Steyer, R., & West, S. G. (2013). Theory and analysis of total, direct, and indirect causal effects. Unpublished manuscript, Psychologisches Institute, Universität Jena, Jena, Germany.
- Mazur-Hart, S. F., & Berman, J. J. (1977). Changing from fault to no-fault divorce: An interrupted time series analysis. *Journal of Applied Social Psychology*, 7, 300–312.
- Ming, K., & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable numbers of controls. *Biometrics*, 56, 118–124.
- Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inferences: Methods and principles for social research*. New York: Cambridge University Press.
- Moser, S. E., West, S. G., & Hughes, J. N. (2012). Trajectories of math and reading achievement in low achieving children in elementary school: Effects of early and later retention in grade. *Journal of Educational Psychology*, 104, 603–621.
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New

York: Oxford University Press.

- Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health*, 94, 423–432.
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd Ed.). New York: Cambridge University Press.
- Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis*, 31, 463–479.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis*. Thousand Oaks, CA: Sage.
- Reichardt, C. S. (2006). The principle of parallelism in the design of studies to estimate treatment effects. *Psychological Methods*, 11, 1–18.
- Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
- Reichardt, C. S., & Mark, M. M. (2004). Quasi-experimentation. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (2nd Ed., pp. 126–149). San Francisco: Jossey-Bass.
- Reis, H. T., & Gosling, S. D. (2010). Social psychological methods outside the laboratory. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., Vol. 1, pp. 82–114). New York: Wiley.
- Reynolds, K. D., & West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review*, 11, 691–714.
- Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. A., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning*, 19, 1–25.
- Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, 7, 331–363.

- Roos, Jr., L. L., Roos, N. P., & Henteleff, P. D. (1978). Assessing the impact of tonsillectomies. *Medical Care*, 16, 502–518.
- Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics*, 11, 207–224.
- Rosenbaum, P. R. (1987). The role of a second control group in an observational study (with discussion). *Statistical Science*, 2, 292–316.
- Rosenbaum, P. R. (2002). *Observational studies* (2nd Ed.). New York: Springer-Verlag.
- Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102, 191–200.
- Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer-Verlag.
- Rosenbaum, P. R., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Rubin, D. B. (1974). *Estimating causal effects of treatments in randomized and nonrandomized studies*. *Journal of Educational Psychology*, 66, 688–701.
- Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2, 1–26.
- Rubin, D. B. (1978). Bayesian inference for causal effects. *The Annals of Statistics*, 6, 34–58.
- Rubin, D. B. (1986). What ifs have causal answers. *Journal of the American Statistical Association*, 81, 961–962.
- Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100, 322–331.
- Rubin, D. B. (2011). Statistical inference for causal effects, with emphasis on applications in psychometrics and education. In M. Williams & P. Vogt (Eds.), *Handbook of innovation in social research methods* (pp. 524–542). Thousand Oaks, CA: Sage.
- Sagarin, B. J., Ratnikov, A., Homan, W. K., Ritchie, T. D., & Hansen, E. J. (in press). Treatment noncompliance in randomized experiments: Statistical approaches and design issues. *Psychological Methods*.

- Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, 13, 279–313.
- Schwarz, N. B., & Hippler, H. J. (1995). Subsequent questions may influence answers to preceding questions in mail surveys. *Public Opinion Quarterly*, 59, 93–97.
- Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51, 515–530.
- Seaver, W. B., & Quarton, R. J. (1976). Regression discontinuity analysis of the dean's list effects. *Journal of Educational Psychology*, 68, 459–465.
- Sechrest, L., West, S. G., Phillips, M. A., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: Strength and integrity of treatments. In L. Sechrest, S. G. West, M. Phillips, R. Redner, & W. Yeatons (Eds.), *Evaluation studies review annual* (Vol. 4, pp. 15–35). Beverly Hills, CA: Sage.
- Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods*, 15, 3–17.
- Shadish, W. R. (2013). Propensity score analysis: Promise, reality, and irrational exuberance. *Journal of Experimental Criminology*, 9, 129–144.
- Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association*, 103, 1334–1343.
- Shadish, W. R., & Cook, T. D. (1999). Design rules: More steps toward a complete theory of quasi-experimentation. *Statistical Science*, 14, 294–300.
- Shadish, W. R., & Cook, T. D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology*, 60, 607–629.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston: Houghton-Mifflin.

- Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., Cook, T. D. (2011). A randomized experiment comparing random to cutoff-based assignment. *Psychological Methods*, 16(2), 179–191.
- Shadish, W. R., Hu, X., Glaser, R. R., Knonacki, R., & Wong, S. (1998). A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes. *Psychological Methods*, 3, 3–22.
- Shadish, W. R., & Sullivan, K. J. (2012). Theories of causation in psychological science. In H. Cooper (Ed.), *APA Handbook of research methods in psychology* (Vol. 1, pp. 23–52). Washington, DC: American Psychological Association.
- Shadish, W. R., Hedges, L. V., Pustejovsky, J., Rindskopf, D. M., Boyajian, J. G. & Sullivan, K. J. (in press). Analyzing single-case designs: *d*, *G*, multilevel models, Bayesian estimators, generalized additive models, and the hopes and fears of researchers about analyses. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-Case Intervention Research: Methodological and Data-Analysis Advances*. Washington, D.C.: American Psychological Association.
- Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd Ed.). Thousand Oaks, CA: Sage.
- Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? *Journal of the American Statistical Association*, 101, 1398–1407.
- Steyer, R., Partchev, I., Kroehne, U., Nagengast, B., & Fiege, C. (in press). *Probability and causality: Theory*. Heidelberg, Germany: Springer.
- Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *The Journal of the Royal Statistical Society, Series A*, 174, 369–386.
- Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, 44, 395–406.
- Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral*

Research, 46, 514–543.

Trochim, W. M. K. (1984). *Research design for program evaluation: The regression-discontinuity approach*. Beverly Hills, CA: Sage.

Trochim, W. M. K., Cappelleri, J. C., & Reichardt, C. S. (1991). Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: II. When an interaction effect is present. *Evaluation Review*, 15, 571–604.

Velicer, W. F., & Molenaar, P. C. (2013). Time series analysis for psychological research. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology, Vol. 2: Research methods in psychology* (2nd Ed., pp. 628–660). Hoboken, NJ: Wiley.

Vinokur, A. D., Price, R. H., & Caplan, R. D. (1991). From field experiments to program implementation: Assessing the potential outcomes of an experimental intervention program for unemployed persons. *American Journal of Community Psychology*, 19, 543–562.

Warner, R. M. (1998). *Spectral analysis of time-series data*. New York: Guilford.

West, S. G. (2008, July). Observational studies: Towards improving design and analysis. In Symposium on causal effects – design and analysis. Altes Schloss Dornburg, Germany. Video available from <http://www.metheval.uni-jena.de/projekte/symposium2008/>

West, S. G. (2009). Alternatives to randomized experiments. *Current Directions in Psychology*, 18, 299–304.

West, S. G., & Aiken, L. S. (1997). Towards understanding individual effects in multiple component prevention programs: Design and analysis strategies. In K. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 167–210). Washington, DC: American Psychological Association.

West, S. G., Aiken, L. S., & Todd, M. (1993). Probing the effects of individual components in multiple component prevention programs. *American Journal of Community Psychology*, 21, 571–605.

West, S. G., Cham, H., Thoemmes, F., Renneberg, B., Schultz, J., & Weiler, M. (in press). Propensity scores as a basis for equating groups: Basic principles

and application in clinical outcome research. *Journal of Consulting and Clinical Psychology*.

- West, S. G., Duan, N., Pequegnat, W., Gaist, P., DesJarlais, D., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M., & Mullen, P. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, 98, 1359–1366.
- West, S. G., & Graziano, W. G. (2012). Basic, applied, and full-cycle social psychology: Enhancing causal generalization and impact. In D. T. Kenrick, N. J. Goldstein, & Braver, S. L. (Eds.), *Six degrees of social influence: Science, application, and the psychology of Bob Cialdini* (pp. 119–133). New York: Oxford University Press.
- West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of Personality*, 59, 611–662.
- West, S. G., Hepworth, J. T., McCall, M. A., & Reich, J. W. (1989). An evaluation of Arizona's July 1992 drunk driving law: Effects on the city of Phoenix. *Journal of Applied Social Psychology*, 19, 1212–1237.
- West, S. G., Newsom, J. T., & Fenaughty, A. M. (1992). Publication trends in *JPSP*: Stability and change in the topics, methods, and theories across two decades. *Personality and Social Psychology Bulletin*, 18, 473–484.
- West, S. G., Ryu, E., Kwok, O-M., & Cham, H. (2011). Multilevel modeling: Current applications in personality research. *Journal of Personality*, 79, 2–49.
- West, S. G., & Sagarin, B. J. (2000). Participation selection and loss in randomized experiments. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (pp. 117–154). Thousand Oaks, CA: Sage.
- West, S. G., & Thoemmes, F. (2008). Equating groups. In J. Brannon, P. Alasuutari, & L. Bickman (Eds.), *Handbook of social research methods* (pp. 414–430). Thousand Oaks, CA: Sage.
- West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods*, 15, 18–37.
- Willett, J. B., & Singer, J. D. (2013). *Applied multilevel data analysis*. Book in preparation. Graduate School of Education, Harvard University, Cambridge, MA.

Wilson, T. D., Aronson, E., & Carlsmith, K. (2010). The art of laboratory experimentation. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th Ed., Vol. 1, pp. 51–81). Hoboken, NJ: Wiley.

Wong, V. C., Steiner, P.M., Cook, T. D. (2013). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics*, 38, 117--141.

¹ In fact, the probability of being assigned to the treatment and control conditions may differ. For example, the probability of assignment to the treatment group might be .25 and the probability of assignment to the control group might be .75 for each participant. Unequal allocation of participants to treatment and control groups is normally used when the cost or difficulty in implementing one of the treatment conditions is substantially greater than for the other. We will assume equal allocation of participants to treatment and control groups here. Given the usual statistical assumptions underlying hypothesis testing, this equal allocation strategy maximizes the power of the test of the null hypothesis.

² Researchers typically only see the causal effect estimate from their own single experiment. They have worked hard designing the experiment, recruiting and running the participants, and analyzing the data and consequently have great confidence in the results of their single experiment. When another researcher fails to find the same result, it is easy to attribute his lack of findings to methodological problems (the “crap research” hypothesis; Hunter, 1996). However, Hunter argues that meta-analyses of several research areas have suggested that methodological quality accounts for relatively little of the variability in the causal effects. Rather, simple sampling variability of the type illustrated here appears to be the main source of the variability in estimates of causal effects (Hunter, 1996).

³ The terminology in the missing data area can be confusing. Missing completely at random means that missingness (whether a variable is observed or not for each participant) is purely random. No adjustment for missingness is needed in the results. Missing at random means that missingness is random after

controlling for measured variables. Following this adjustment, the ACE will be an unbiased estimate of the causal effect in the population. Missing not at random means that missingness depends on the unobserved values of the missing variables and so that an unbiased estimate of the ACE may not be possible.

⁴ The approach of throwing out participants has been the standard procedure in laboratory experiments in social psychology when participants are suspicious, uncooperative, or when other problems arise. The possibility that this procedure introduces potential bias should always be considered.

⁵ In randomized field experiments, participants are nearly always more similar within than between groups. In other situations involving within-subject designs or other forms of dependency, the direction of bias may change (see Kenny & Judd, 1986).

⁶ Although less often hypothesized by social psychologists, changes in the variance of the series or in cyclical patterns in the series following an intervention can also be detected. Larsen (1989) and Warner (1998) have discussed ways in which cyclical patterns of variables such as mood and activity levels may be important in human social behavior.

⁷ If important baseline differences remain after propensity score matching, the propensity scores can be re-estimated using another model (e.g., adding interactions). Alternatively, important covariates on which differences remain can be statistically controlled in the main outcome analysis. Rosenbaum (1987, 2010) also offers methods of conducting sensitivity analyses that probe how large an effect would be needed on an unmeasured covariate to alter the results of the main outcome analysis.

Chapter five Field Research Methods

Elizabeth Levy Paluck and Robert B. Cialdini

Introduction

“In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that that vast Map was Useless.”

Jorge Luis Borges, “On exactitude in science,” 1999

Theories are like maps. Social psychologists typically use theories as maps to find phenomena that are worth investigating, or to deduce unexplored pathways between those phenomena. After formulating their hypotheses, they design a laboratory experiment that simulates the local conditions described in their theoretical map.

Social psychologists often do not see a place for field research in this cycle between theory and laboratory experimentation. Field research, they fear, will bloat their theoretical maps with too many added variables, distorting causal pathways. Observational fieldwork used to identify interesting phenomena, or experimentation outside of the highly controlled laboratory environment, so the argument goes, will produce unwieldy theoretical maps akin to Borges's maps that were the exact size and scale of the Empire.

Years ago, one of us (Cialdini, [1980](#)) pointed out that on the contrary, field research can help social psychologists draw accurate theoretical maps that identify the most consequential social psychological phenomena. While theoretically driven laboratory experimentation can produce accurate maps, they may not tell social psychologists about the most interesting or important

locations. Furthermore, it is by cycling through field observation, experimentation, and theory that social psychological theories can become precise as well as meaningful. This claim is rooted in a long-standing call for more fieldwork in social psychology (Campbell, 1969; Lewin, 1947; McGuire, 1969), and in Cialdini's own “full cycle” research program, which used field observation, field experiments, laboratory experiments, and theoretical deduction to develop a framework for social influence (Cialdini, 2009a).¹

This chapter will not cover the uses of field research for the *application* of social psychological theory. Instead, this chapter describes the use of field research for *development* of psychological theory. Field research fosters theoretical development in many ways. Field research helps identify which phenomena are most psychologically and behaviorally consequential. Operationalizing independent and dependent variables and choosing the right setting in the field compels researchers to specify and make concrete their theoretical constructs of interest. Field research allows investigators to observe some possible boundary conditions of their theory and to examine how different theoretical constructs relate to one another. The field is also a good setting for testing causal predictions. Of course, theory that has been developed from field research stands a better chance of successful application to real-world issues. However, in this chapter, we will focus on the kinds of theoretical insights afforded by research in field settings.

Most students of social psychology are drawn to the discipline because of an interest in the world around them, but in the course of study, their eyes are retrained to find inspiration in abstract theory and to observe and test these theoretical processes in laboratories. Many important figures in social psychology, themselves experts in laboratory experiments (e.g., McGuire 1969), have lamented this rather myopic methodological focus. There are a few reasons why field-based observation, measurement, and experimentation have not historically been as prominent as laboratory work (Cialdini, 2009b). One simple reason is that social psychologists are not typically trained as a matter of course in field methodology.

We have designed this chapter to be a systematic treatment of various options in field research, so as to redirect students' and researchers' eyes toward these methods. It is our hope that awareness of the uses and advantages of these field methods, paired with an understanding of when and how they can be implemented, will promote more social psychological research in the field.

The chapter is laid out as follows. We first explain what we mean by field

research as opposed to laboratory research, and discuss advantages that come from finding and testing ideas in the field. We explore the range of theoretical goals that can be accomplished with field research. We point out strengths and weaknesses of various field research techniques and some best practices of each one. We conclude with practical suggestions and reasons for researchers at various stages of experience to engage in field research.

Laboratory and Field Research

What Is Field Research?

“Field” research is, of course, not defined by its physical locale, but by the work's degree of naturalism. Defining field research as relatively more naturalistic elements allows for a continuum-based (rather than dichotomous) conceptualization of the approach. After all, the laboratory can be the site of very realistic interventions, and conversely, artificial interventions may be tested in a non-laboratory setting. When assessing the degree to which studies qualify as field studies, one must consider the naturalism of four aspects of the study: (1) participants, (2) the intervention and its target, (3) the obtrusiveness of intervention delivery, and (4) the assessed response to the intervention.

For example, a study on the effects of interpersonal empathy might involve (a) undergraduate psychology majors, (b) written instructions aimed at the participant's perception of a sad story, which vary systematically according to whether (c) instructions to empathize with the protagonist of the story are included or not, and (d) outcome measures such as empathy and willingness to help scales. Relatively more naturalistic versions of each aspect of this study are (a) non-psychology major young adults or citizens of the local town, (b) a television station broadcasting a sad story, which varies systematically in terms of its (c) language and imagery that encourages or does not encourage the participant to take the perspective of the protagonist, and (d) measures such as the participant's facial expression as they watch the screen, or their response to a nearby confederate who disparages the protagonist. Note that this experiment could be conducted in a laboratory that has been outfitted to look like a waiting room with a television, which would make the laboratory more naturalistic.

Cronbach (1982) suggests an acronym, UTOS, to use when assessing the naturalism of a study: Units, Treatments, Observations, and Setting. To this list of considerations, Reis and Gosling (2010) add that non-laboratory research

settings differ from the laboratory in the goals that are likely to be activated by the setting, the setting's correspondence with the behavior under study, and the degree to which the setting is natural and appropriate for the research question.

Advantages of the Laboratory

Before we detail some of the disadvantages of laboratory settings that are addressed through research in field settings, it is important to recognize the many advantages of the laboratory for social psychological research. The laboratory is singular for its precision and control, which produces low error variance and nuanced and well-targeted measurement. Control over variations in treatment allows the investigator to fully stage-manage and test the interaction of the setting with participants' individual differences. Investigators are free to eliminate or include any variables they determine to be extraneous or potentially influential. In this way, investigators can test theory from all angles, probing mechanisms and counterintuitive predictions of the guiding theory (Smith, [Chapter 3](#) in this volume; Wilson, Aronson, & Carlsmith, [2010](#)).

Laboratory research is also convenient for university investigators. Laboratories can be located next door to the investigator's office, for easy supervision of research assistants and the research site itself. Undergraduate psychology majors are efficiently exposed to psychological research in on-campus laboratories, and critically, they serve as participants in laboratory research so that investigators can execute multiple studies per semester.

Disadvantages of the Laboratory

What is the cost of the predominance of the laboratory research in social psychology? It is theory's "close relation to life," according to Kurt Lewin, one of the founders of modern social psychology (Lewin, [1944/1997](#), p. 288). Relating back to the idea of theories as maps, Cialdini ([1980](#)) wrote that theory-driven experimentation without attention to the real world could result in an accurate but less consequential map, or even a misleading map. As is often the case, the greatest strength of laboratory research – its control – is also part of its weakness. We elaborate on this point later in the chapter in terms of Cronbach's (1982) scheme of Units, Treatments, Observations, and Settings, and in terms of psychologists' concerns about culture, complex systems, and identification of "extraneous" influences.

Units. Undergraduate students are predominantly used as participants in

psychological laboratory settings because of their convenience, tradition, and financial discount for academic investigators. For the purposes of building widely applicable theories, undergraduates present several troubling bias. Their developmental stage and their particular social and educational backgrounds may exaggerate some effects and diminish others, or restrict the range of variation on the dimensions being studied (Henry, 2008; Sears, 1986).

Treatments. The treatments administered in laboratory settings are typically weaker, briefer, and less varied than the naturally occurring phenomena in the world that they are designed to simulate. The fact that treatments are weaker is often due to experimenters' ethical obligation to avoid intense or distressing events, such as authority coercion, severe disappointment or sadness, or sexual harassment (cf. LaFrance & Woodzicka, 2005).

Laboratory treatments are brief to accommodate the typical hour-long sessions allotted to participants. As a consequence, researchers use a "reactive or acute form" of a variable to stand in for longer-term phenomena. For example, when studying low self-esteem, laboratory experimenters must lower self-esteem with negative feedback or an experience of failure, rather than observe the process of erosion of self-esteem over a longer term. Unfortunately, "an occasion of low self-esteem may have nothing to do with a lifetime of low self-esteem" (Ellsworth, 1977, p. 607). Short-term states may not operate under the same underlying processes as chronic states, which are almost impossible to study experimentally in the laboratory over long time periods (cf. Cook's [1978] months-long laboratory studies of interracial workgroups).

Finally, there is often little variety in the types of treatments used in the laboratory. Investigators rely on a few established paradigms to study a variety of outcomes, and very rarely translate their abstract theoretical ideas into new operationalizations. "The mental dexterity demonstrated in dealing with abstractions often seems to vanish at the translation stage, as the old standard treatments and measures are used and reused with little consideration of their suitability for the task at hand (i.e., choosing a concrete version of an abstract question)" (Ellsworth, 1977, p. 604; see also Webb, Campbell, Schwartz, & Sechrest, 1966).

Observations. Very rarely are the behaviors measured in the laboratory commensurate with the behaviors that investigators wish to explain in the real world. First, many outcomes examined in the lab, such as reaction times, are rarely important outcomes in and of themselves in real-world settings. Second, lab-based pseudo-behaviors, such as deciding the salary of a fictional person or

assigning a sentence to a fictional criminal in a jury vignette, may not result from the same interpersonal and intrapersonal processes that produce these behaviors in the world. This difference is troubling for theory testing and building. Most often, laboratory investigations measure self-report rather than behavior (Baumeister, Vohs, & Funder, 2007).

Settings. Social psychologists strive to study the interaction of the person and the situation, but there is very little work that describes situations themselves (cf. Kelley, Holmes, Kerr, Reis, Rusbult, & Lange, 2003; Reis, 2008) or innovates different situational paradigms in the laboratory. As a result, investigators cannot observe and catalog the situations that are most frequent or consequential for individual or group behavior.

The culture of the laboratory. One presumed advantage of the laboratory is that it is a “culture-free” setting – one that is not tied to any particularistic traditions, scripts, ideologies, or standards of reference. Adams and Stocks (2008) argue that this assumption is misguided, and has given rise to theories that are incorrectly portrayed as universal processes of human cognition and behavior. Examples of cultural elements of the laboratory are Likert scales featuring implicit standards of reference that are culturally specific (e.g., comparing oneself to another individual or to another group; Heine, Lehman, Peng, & Greenholtz, 2002, cited in Adams & Stocks, 2008), and exercises that rely on culturally specific ideas of relationships (e.g., trust-building exercises that involve self-revealing information; Aron, Melinat, Aron, Vaollone, & Bator, 1997).

Related to these points, investigators have found that participants harbor social scripts and expectations for laboratory situations that affect their behavior and thus threaten the external validity of the research (Shulman & Berman, 1975). For example, Bator and Cialdini (2006) argue that certain features of the laboratory as well as its scientific atmosphere stimulate research participants to respond in more logically consistent ways than outside of the laboratory. Other researchers worry that investigators rarely implement methodological solutions to prevent artifacts such as experimenter bias, subject motivation, and meta-processing of the situation by participants (West & Graziano, 2012).

Complex systems in the laboratory. The laboratory is often an inappropriate setting for studying complex systems, which is troublesome given the complexity of human behavior. “More and more, we are coming to recognize that [variables’] interrelations may be causal but much more complicated than we can assess with our usual methods” Ellsworth (1977, p. 614) asserts. “It is in

just these instances [of complex relationships] that the typical laboratory experiment is weakest; so much is held constant that there is no opportunity for this sort of complex causation to manifest itself.”

Moreover, there are many unobservable variables operating in a real-world context, variables of which investigators may not be aware when they set out to simulate that context in the laboratory. Consequently, a relationship uncovered between two variables in the laboratory may not exist or may occur rarely in the world because of the interference of this unobserved variable. Preceding, co-occurring, or proceeding variables in the real world may diminish the relationship identified.

While we have cataloged many disadvantages of laboratory settings and research paradigms, these critiques should not be taken solely as arguments to incorporate fieldwork into a research program. These preceding points can also be used to inspire more rigorous laboratory experiments that test and produce theoretical maps that are, in Lewin's words, “closer to life.”

Advantages of the Field

The most obvious advantage of the field is that the investigator does not always have to work as hard to make the units, treatments, observations, or settings of a study naturalistic. Participants are those people involved in the treatment or who come from the social group of interest; treatments can be more high impact and lengthy than a laboratory intervention; outcome measures can be those that already occur in that setting.

Definition of constructs. Selecting the location and the participants for a field-based study helps investigators define precisely the nature and scope of their theoretical constructs. Take the following example. Suppose that you are interested in studying cooperation. You understand that your choice of setting (e.g., households of married couples, a cheese cooperative in Berkeley, a financial trading floor) and your participant population (e.g., adults, kindergarteners, residents of a small-scale agriculturalist society) change what you mean by “cooperation” and how you will measure it. As you eliminate certain types of settings and populations, you refine your concept of what kind of cooperation you will be able to describe and theoretically map in relationship to other constructs. Leaving the laboratory's standardized paradigms and generating a list of possible settings, participants, and measurements reveal implicit assumptions or theoretical confusions about your construct (Ellsworth, [1977](#)).

Inductive power. Another advantage of research in field settings is that it can provide an inductive approach to theory that begins with facts about cognition, emotion, and behavior in the world, rather than a deductive approach that begins with abstract theory. Cialdini (1980) described the inductive capacity of field research as a “steadily developing sense of which of our formulations account not just for aspects of human behavior, but also for aspects of the behaviors that matter” (p. 26). Moreover, to generate ideas in the first place, McGuire (1969) suggested that investigators would spend their time more productively observing field settings rather than reading the top journals. Using fieldwork to establish the strength, frequency, and surrounding conditions of an effect is a powerful approach to assembling the building blocks of a new theory or to modifying an existing theory.

Causal testing. As we discuss later, the field is not just a setting for observational research. Field settings provide a powerful stage for causal tests. Just as laboratory experimentalists use stagecraft to import various conditions of the real world into the laboratory, field experimentalists export experimental control from the laboratory into the field. Testing causal relationships in the field allows investigators to identify whether the relationships hold up in the presence of other social and situational factors. Field experiments also indicate plausible boundary conditions of an effect across different time periods, settings, varying numbers of people, and other important contextual factors.

Test of a theory's pragmatic worth. Saying that field experiments reveal whether causal relationships hold up in real-world settings is one way of saying that field experiments test the *pragmatic* worth of a theory. By pragmatic we do not simply mean applied. We use pragmatic in the way William James (1981) defined pragmatism, specifically that theories are pragmatic when their predictions “cash out” in the real world – when they predict behavior occurring in the “rich thicket of reality” (p. 517; see also Brewer & Crano, Chapter 2 in this volume; Fiske, 1992). Field settings invite psychologists to be concerned not just about their stock in the marketplace of ideas, but also about their stock in the marketplace of observable effects.

Relevance. Relatedly, field research renders social psychological theory and research more valuable to members of important nonacademic communities. Because it takes place in natural, everyday settings, field research makes transparent the relationship between the obtained data and everyday lives. That clear relevance allows those who have paid for the work (e.g. taxpayers and research purse-holders) and those who would want to employ it (e.g., policy and

decision makers) to view social psychologists as a credible source of information about the issues of concern to them. Some evidence that social psychologists have yet to be viewed in this way by certain important individuals comes from a pair of experiences of one of the authors. At two separate meetings of high-level government officials, he was labeled not as a social psychologist but as a behavioral economist because, he was informed, it was judged to be more palatable to the participants.

Field Observational Methods

Observation in the field is not a supplement to empirical work – it is empirical work. For example, Cialdini's (1980) full-cycle model endorses “a more empirical science that is based firmly in the observation of everyday worlds” (Adams & Stocks, 2008, p. 1900). From this perspective, observation and experimentation are each “but one tool in the social psychologist's repertoire,” and each is “better suited to some tasks than others” (Adams & Stocks, 2008, p. 1900).

Observational methods can be put to many important uses in field settings. Observation of individual and group behavior can generate hypotheses and theoretical insights, or point researchers toward phenomena that are powerful and prevalent in the community (Mortensen & Cialdini, 2010). Or, as Solomon Asch once pointed out, observation can help researchers become more familiar with the phenomenon of interest: “Before we inquire into origins and functional relations, it is necessary to know the thing we are trying to explain” (Asch 1952, p. 65, cited in Reis and Gosling, 2010). This includes observing and describing the types of situations that give rise to the phenomenon (Kelley et al., 2002). In addition, observational measurement techniques like interviews or behavioral trace indicators, described later in the chapter, can be used as outcome measurements for field or laboratory experiments. All observational measures, particularly those that are highly unobtrusive, can serve as strong verification of self-report data.

Qualitative Methods

Qualitative methods used to explore or describe phenomena include personal observation, participant observation, structured interviews, and ethnography. Notes produced by these methods can be coded and written up for publication (Emerson, Fretz, & Shaw, 1995). If the data were collected in a systematic

manner, qualitative outcomes may be used as outcome measurements in a study by quantitatively coding and analyzing the data as events or ratings (Paluck, 2010).

Personal observation. Personal observation is a time-honored tradition of hypothesis generation in social psychology. A classic example is Festinger's (1957) observation that catastrophes are followed by rumors of further disaster rather than reassurances of relief, which led to his formulation of cognitive dissonance theory. Although many social psychologists report that ideas and counterintuitive notions were inspired by real-world observation, observational skills are not often recognized as part of the social psychologist's official toolkit. McGuire (1973) urged psychologists to “[cultivate] habits of observation that focus one's attention on fertile aspects of natural experience,” adding “[we should] restructure our graduate programs somewhat to keep the novice's eye on the real rather than distracting and obscuring his view behind a wall of data” (p. 453). Keeping one's “eye on the real” could involve training oneself to be more alert in everyday life, delving into written accounts of everyday life through various peoples’ eyes in blogs or newspapers, or using more systematic observation techniques like participant observation.

Participant observation. Participant observation involves observation while participating in an institution, social group, or activity. For example, investigators can participate in skilled practitioner trainings, as did Cialdini (1993) in his observation of sales trainings. These observations helped him formulate important underlying principles of compliance tactics salespeople had honed over years of work. Some psychology departments send students out to participate in community organizations, to observe and analyze where social psychology can contribute (Linder, Reich, & Braver, 2012). “Being on the scene often means a necessary exposure to a large body of irrelevant information,” Webb, Campbell, Schwartz, Sechrest, and Grove (1981, p. 240) caution, but “the payoff is often high.”

Ethnography. Ethnographers spend concentrated amounts of time in a particular place or following a particular group of people or event (e.g., a neighborhood, or a traveling political rally). Rather than seeking to measure the frequency of a behavior in a setting, ethnographic methods are aimed at understanding the social psychological *meaning* of that behavior in the context. And rather than collecting a representative sample of people or places, ethnographers focus on one or a few “cases,” such as individuals, classrooms, teams, or towns. Ethnographers try to get to know their subjects and to become

part of their lives and contexts for a period of time (Lareau & Schultz, [1996](#)). For example, Erving Goffman wrote the foundational text, *The Presentation of Self in Everyday Life* (1959), after one year of living in and observing a Shetland Island subsistence farming community.

Interviews. Field research can also benefit from structured or in-depth interviews with individuals, called key informants, who have specialized experience with the phenomenon or community under investigation. For example, Huggins, Haritos-Fatouros, and Zimbardo ([2002](#)) interviewed Brazilian police who had tortured and killed citizens during Brazil's military rule, to understand the process by which they were convinced to commit atrocities on behalf of the state, and how they justified this violence to themselves and their peers. Adams ([2005](#)) combined interviews with field observation to uncover the concept of enemyship (a personal relationship of hatred) in West Africa and North America.

Observation-Based Estimates of Individual or Population Characteristics

Individual and population characteristics can be inferred from observational field methods such as daily diary techniques, trace measures, ambulatory assessment, and social network mapping (see also Reis, Gable, & Maniaci, [Chapter 15](#) in this volume). These types of observational data can be collected in person, in archives, or can be harvested from the Internet.

Individual characteristics. Daily diary methods “capture life as it is lived” (Bolger, Davis, & Rafaeli, [2003](#), p. 95). Participant are asked to fill out reports about their behavior, affect, cognition, and/or surroundings at regular intervals or when prompted at random times by a PDA or a mobile phone. Diary methods serve the important descriptive purpose of cataloging information about the prevalence, chronological timing, and co-occurrence of events and situations (Reis & Gosling, [2010](#)).

Trace measures bring out the Sherlock Holmes in social psychologists. To track psychologically meaningful behavior unobtrusively, psychologists seek out systematically or automatically recorded traces of behavior in official archives or unofficial spaces of everyday life. The advantage of these measures is that the subjects of study are unaware that they are being watched. For example, from official public records, investigators can study government voting or hospital immunization records and yearbook photos. Investigators might even obtain data

from retail stores on customer loyalty card activity showing individuals' purchases of fruits and vegetables, cigarettes, or other products. From the "unofficial" records, social psychologists have mined trash cans to measure alcohol consumption (Rathje & Hughes, 1975) and counted the number of vinyl tiles that needed to be replaced around various museum exhibits as a measure of interest in the exhibit (Duncan, personal communication, cited in Webb et al., 1981). Analyzing the composition of personal Internet profiles on social networking websites is one of the latest ways to use trace measures (Reis & Gosling, 2010). Online social networking sites also help investigators to identify individual's network of potential social influences (e.g., Goel, Mason, & Watts, 2010).

Another individual-level measurement technique is ambulatory assessment, which uses electronic equipment to measure an individual's movement and states of being throughout their daily lives. This includes blood pressure monitors, sound recording (Pennebaker, Mehl, Crow, Dabbs, & Price, 2001), and GPS tracking devices located in individuals' mobile phones. Some tracking devices can even assess which people in a social network, such as a school social network, interact the most frequently, and for how long.

The advantage of these observational methods is that they capture daily experiences as they occur in the stream of natural activity across different situations. Many of these methods produce time series data, which means they can evaluate hypotheses regarding the chronological ordering of a particular process and within-person processes (Reis & Gosling, 2010, pp. 96–97). Depending on the way they are collected, observational measures can overcome biases of self-report. For example, in the garbage trace measures collected by Rathje and Hughes (1975), 15% of households reported at the front door that they drink beer, while beer cans were found in the trash can at the back door in 77% of the same group of households.

Disadvantages of individual observational methods include noncompliance or misreporting, in the case of daily diaries. In addition, there may be imprecise translation between trace measures and psychological constructs or behavior. For example, did individuals actually drink the beer from the cans found at the back door, or did they use the beer to bake bread? This may be an overly generous interpretation of their trash cans, but for observational measures the general principle holds that measures are strongest when they are deployed alongside different types of measures that can corroborate their findings.

Population characteristics. Some observational methods cannot connect

observations to specific individuals, but can draw a picture of a community as a whole. For example, a linguistic analysis of online journals before, during, and after the events of 9/11 revealed average social psychological reactions to trauma among U.S. residents (Cohn, Mehl, & Pennebaker, 2004). The lost-letter technique is another extensively used population-level observational method. Throughout the streets of a community, investigators drop stamped, addressed letters (while the name varies, the address sends the letter to the investigator). The proportion of letters that are picked up and delivered to a mailbox serves as a measure of average community helpfulness. Investigators have extended the purpose of this technique to measuring social bias. The names listed on the letter's address are randomized such that half feature typically Anglo-American names and the other half African-American names. Investigators measure whether the proportion of redelivered letters differs depending on the presumed race of the recipient (Crosby, Bromley, & Saxe, 1980).

Population-level observation is the most unobtrusive of the research methods reviewed here, avoiding completely the “speak clearly into the microphone, please” aspect of other approaches (Webb et al., 1981, p. 241). In this sense, such observation holds an advantage over laboratory settings in which participants know that their behaviors are being examined, even if they do not know which behaviors those are. Population-level observations do, however, prevent the investigator from connecting individuals to behaviors, which means that these outcome measures are best used for description, hypothesis generation, or for an experiment in which the community is the unit of randomization and analysis.

Observation of Situation Characteristics

Despite the stated importance of the situation in social psychological analysis (Rozin, 2001), very little observational work has been devoted to establishing a taxonomy of different situations. The work of Kelley *et al.* (2003) is a notable exception, in which the authors classify and describe 21 of the most common everyday situations thought to influence various aspects of interpersonal behavior. We believe that psychologists could make greater use of this work and expand on it with situational taxonomies relevant to other types of behavior, cognition, or emotion.

Ultimately, observational research in field settings involves detecting “pieces of data not specifically produced for the purpose of comparison and inference but available to be exploited opportunistically by the alert investigator” (Webb et

al., 1981, p. 5). We now turn to experimental research in the field that is explicitly designed for the purpose of comparison and causal inference.

Field Experimental Methods

Experimentation in field settings can be just as rigorous as in a laboratory setting. Treatments can be randomly assigned and delivered in a standardized manner to individuals, groups, or institutions, and standardized outcome measures and evidence speaking to the process of change can be collected. Causal inference in field settings has greatly improved over the years, mostly through innovations in field experimental design that address challenges particular to field settings (Green & Gerber, 2012; Rubin, 2005; Shadish, Cook, & Campbell, 2002; West, Cham, & Liu, Chapter 4 in this volume).

Randomization and Control in Field Settings

Psychologists who conduct laboratory experiments may approach field experimentation with two types of reservations. One concern is that ongoing activity in field settings will destroy pure randomization and segregation of experimental and control groups. A second is that many things are simply impossible to randomize in a field setting. On the first point, psychologists might be pleasantly surprised to find the many varieties of experimental designs field experimentalists have developed to preserve the integrity of experimental design against special situations that arise in the field.

On the second point, canvassing the types of psychological field experiments conducted over the last decade reveals few limits on the kinds of treatments that have been randomized. Social psychologists have randomized a variety of interventions, in national parks (Cialdini, 2003), on national radio (Paluck, 2009), in schools (Blackwell, Trzesniewski, & Dweck, 2007), and amusement parks (Gneezy, Gneezy, Nelson, & Brow, 2010), targeting and measuring psychological phenomena from perceived norms, beliefs and implicit theories to social welfare concerns, and connecting them with real-world behavior. For most social psychologists, whether or not their interests lie in basic or applied research, these are important and worthy investigations.

Many field experiments involve simple random assignment of a treatment to individuals, households, or communities. Investigators deliver the treatment, or they collaborate with an organization that is already intending to deliver the treatment. However, many types of treatments are impossible to package neatly

and randomly deliver to some individuals but not others. The following types of designs address some of the issues that arise for these types of experimental treatments.

Encouragement Designs

One advantage of experimenting in field settings is the opportunity to study interventions that are very difficult to simulate in the laboratory, such as political movements. Political movements, however, exemplify the type of treatment that at first blush seems impossible to study experimentally. In the specific case of political movements, the “treatment” is broadcast to the general public, meaning there is no obvious control group. Moreover, joining a political movement is a highly individualized and rare decision, meaning the “treatment” group that joins a political movement is self-selected and small. One experimental design that can capture a treatment with these characteristics is a randomized encouragement design. An encouragement design randomly encourages some people and not others to engage with the treatment and then measures reactions within the entire sample of encouraged versus not-encouraged people.

Consider the following field experimental design to study the effect of joining a political movement on, for example, individual political perceptions and communal behavior. Suppose you identify an organization that has mounted a website calling for political change in a particular city. To measure the causal effects of joining this movement, you could randomly divide a list of city residents’ email addresses in half, and send an email to one-half of the sample. The email would encourage the recipients to visit the site and join the organizers’ efforts. In your entire study population, there will be people who would have visited and joined without encouragement, people who will visit and join despite the fact that they were not encouraged, and people who will never visit or join regardless of encouragement. However, there is also potentially a group of residents who would *not* have visited or joined without encouragement. An encouragement design measures the causal effect of invitation on visitation, participation, and on their subsequent political perceptions and behavior among the encouragement group, compared to the equivalent types of people in the no-encouragement group (for an explanation of analysis of encouragement designs, see Green & Gerber, 2012; Angrist & Krueger, 2001). Because obtaining informed consent is not a typical component of such designs, researchers need to ensure that the invitation and the encouraged activity would not violate participants’ privacy or well-being.

Randomized Rollout Designs

Some high-impact and theoretically relevant treatments are administered by governments or private companies who do not wish to exclude treatment recipients in the interest of forming a control group. In such cases, investigators can use randomized rollout designs to study the causal impact of these treatments. A randomized rollout eventually assigns the entire population to treatment, but over a certain time period, during which outcome measurements can be assessed among the treated and as-yet-untreated participants.

For example, a company that is struggling to attract a diverse workforce may be anxious to implement new hiring accountability measures or a set of diversity trainings. A strategic field experimentalist could explain to the company that, given that an immediate implementation of diversity initiatives to *all* of the company's offices may not be possible for financial or scheduling reasons, a lottery would represent a fair procedure by which to “roll out” the new diversity training. Half of all randomly selected offices could receive the diversity training in the first year and the remaining half in the second. This kind of randomized rollout (or waiting-list design) is ethical in addition to being practical because it allows the company to assess halfway through its implementation whether or not the intervention is having the intended effect (Campbell, 1969; see also Shadish et al., 2002).

Downstream Field Experimentation

One exciting opportunity that is born of a field experiment is downstream field experimentation, or analysis of the indirect effects of an experimental intervention. Policy experiments randomize high-impact treatments that can be expected to set off a chain of events, for example, educational opportunities to low-income students. Investigators can re-contact or gather publically available data on treatment and control students down the road to ask important theoretical questions, for example, whether more education (attained through the college scholarship) changes a person's political ideology, their social values, or the ways in which they raise *their* children (Sondheimer & Green, 2010). Measurement of those outcomes will still represent a causal chain of effect because the educational opportunity itself was randomly assigned. Downstream effects created by preexisting experiments are low-hanging fruit that can be gathered up by graduate students or other investigators with fewer resources.

Hybrid Lab-Field Experiments

As already discussed, some field experiments are more naturalistic than others. Hybrid lab-field experiments are experiments in which elements of artificiality are introduced for purposes of better control over treatment assignment or delivery, or for more precise measurement. Hybrid models are useful when investigators are studying a high-impact independent or dependent variable.

For example, in an experiment on media and interpersonal influence, investigators randomly selected groups of friends in neighborhoods of Juba, South Sudan, to listen for a few hours to a recording of a previously broadcast radio program (Paluck, Blair, & Vexler, 2011). The participants in this study were the target audience of the radio program, and they listened in their own neighborhood with their typical listening partners. The artificial elements of this study were the researchers who sat with the group as they listened, to take notes on group reactions and to interview each group member separately when the program was finished. The laboratory-like surveillance involved in this study detracts from the overall naturalism of the field experiment, but it allows the investigators to obtain precise measurements of attention, verbal and nonverbal communication among the friends, and individual reactions to the program.

Another type of hybrid lab-field experiment is one in which an intervention is delivered in the laboratory and outcome measures are gathered in the field, such as when Walton and Cohen (2007) treated a random half of their sample of minority university students to a belongingness intervention in a laboratory and then followed all of the sampled students' grades over the course of the year. University grades are a high-impact dependent variable, which persuades us of the power of Walton and Cohen's laboratory-based intervention.

Designs to Address Challenges in the Field

Spillover refers to the problem when the treatment or treated participants influence untreated or control participants. For example, a random subset of an apartment building's residents who receive a message encouraging recycling may communicate those messages in passing conversation to the untreated control residents of the building. Attrition refers to the problem of participants dropping out of an experiment, or missing dependent measures for some participants. Attrition is especially problematic if it is differentially triggered by one of the experimental conditions. For example, in an educational experiment, students might drop out of a solitary studying condition more frequently than a

social studying condition. Spillover and attrition are two problems that arise more frequently in field settings compared to laboratory settings.

Because certain types of spillover can underestimate the true effect of the treatment, and more importantly because standard statistical analyses assume that units of a randomized experiment are independent (see West et al., [Chapter 4](#) in this volume), many field experimental designs are set up to prevent spillover between units. The underlying principle of these designs is to select units for your experiment (people, situations, or communities) that are spaced out geographically, or to space your randomly assigned treatments temporally. Alternatively, when it is too difficult to prevent participants from interacting, you can group them together, randomly assigning *clusters* of interacting participants to treatment versus control. Of course, some experiments are explicitly interested in spillover effects, such as the spread of influence throughout a network, and so they use designs that can detect influence among units (see Green & Gerber, 2012, [Chapter 8](#)).

Attrition happens more often in field settings because researchers are more likely to lose track of participants – participants in their natural environments feel less obliged to comply compared to those in laboratories – and because outcome measures are unavailable or blocked by an intervening agent. Of course, attrition is sometimes interesting in itself to study because it may reveal whether your treatment is viable for real-world use. For the most part, however, attrition is an impediment to learning about your experimental effect.

Some statistical approaches to this problem are to make strong assumptions about the potential outcomes among those who dropped out of the study, to put larger bounds around the findings, or to launch a new data collection that attempts to fill in missing values for a randomly chosen subset of the missing participants or outcome measures (Green & Gerber, 2012, [Chapter 7](#)). Increasingly, investigators use technology to minimize attrition in the field, for example, by sending participants text-message reminders to their mobile phones (e.g., Tomlinson, Rotheram-Borus, Doherty, Swendeman, Tsai, le Roux, Jackson, Chopra, Steward, & Ijumba, [2011](#)).

Quasi-Experimentation in the Field

Randomization is always recommended for causal inference in the field, because observational studies will bias average treatment effects (Gerber, Green, & Kaplan, [2004](#); West & Graziano, [2012](#)). However, when random assignment is

not possible in the field, social psychologists have at their disposal many creative designs based on the principles of random assignment and causal inference (Shadish et al., [2002](#)). We mention two of the most prominent designs here.

Regression Discontinuity

A regression discontinuity design is useful when there is no randomization but there is a clear decision point along a continuous measure of eligibility for treatment regarding who will receive the treatment and who will not. If the decision point regarding eligibility is monotonically measured and rigid (i.e., it is monotonically increasing, not nominal like ethnicity, and there are no exceptions to the cutoff), then people who fall just above and just below the decision cutoff are likely to be, on average, comparable. This expectation of average comparability is similar to but not as strong as the expectation of comparability between two groups formed by random assignment. Thus, for the sample of people whose scores fall around the cutoff, investigators can test causal inferences about the treatment received by those who qualified (see Shadish, Galindo, Wong, Steiner, & Cook, [2011](#); West et al., [Chapter 4](#) in this volume).

Interrupted Time Series Analysis

Interrupted time series analysis is used to assess the impact of a treatment that occurs at an observed point during a sustained time period of consecutive observations made on one or a set of outcome measures. Thus, unlike the other designs covered thus far that track outcomes for comparison samples, this design follows one sample over time. A causal relationship between the treatment and the outcome measures is proposed when investigators show that the slope or level of the outcome measures was significantly changed after the treatment. The causal case is strengthened when investigators show that the slope or level of other continuously collected measures unrelated to the treatment were in fact unchanged when the treatment occurred. One example of this method is the study of online journaling before, during, and after 9/11, which shows a change in the way people keeping regular journals responded to trauma as a result of 9/11 (Cohn, Mehl, & Pennebaker, [2004](#)).

The Internet as a Site for Experimentation

As psychologists come to agree that the Internet can be a site of meaningful

social expression, interaction, and behavior, they have profited from Internet sites and samples for psychological experimentation and measurement (Gosling & Johnson, 2010). The Internet is an efficient way to conduct survey experiments that can, depending on the goal, deliver both representative samples of large populations (Berinsky, Huber, & Lenz, 2010) and selective samples of difficult-to-reach populations (e.g., Glaser, Dixit, & Green, 2002). Maniaci and Rogge (Chapter 17 in this volume) provide an extensive treatment of Internet experimentation.

Table 5.1 provides a non-exhaustive list of field research methods and their advantages and disadvantages for psychological research.

Table 5.1. A Non-Exhaustive List of Field Research Methods and Their Advantages and Disadvantages for Psychological Research

Field Research Method	Advantages	Disadvantages
Observational		
Personal observation	Hypothesis generation; observation of strength and frequency of phenomena, of groups, contexts	Difficult to test hypotheses; large amount of qualitative data, some of which may be irrelevant
Participant observation	Observation of a phenomenon, group, or intervention from personal perspective as participant; observation of underlying principles of successful or regularly occurring phenomena in the real	Difficult to test hypotheses; large amount of qualitative data, some of which may be irrelevant; presence of researcher may be obtrusive

world

Ethnography	Observation of many aspects of a phenomenon, group, or context over a longer period of time, aimed at understanding the social psychological meaning of that behavior in the context; researcher may seem less obtrusive over time	Difficult to test hypotheses; large amount of qualitative data to code and analyze, some of which may be irrelevant
Interviews	Access to perspectives of individuals with specialized experience with the phenomenon or community under investigation	Subject to participants' self-report bias and to bias in selection of interviewees
Daily diary	“Capture life as it is lived”; catalog information about the prevalence, chronological timing, and co-occurrence of events and situations	Subject to participants' self-report bias, to bias resulting from selective participant attrition from regular reporting, and to sampling bias from differential willingness to participate; highly obtrusive
Trace measures	Highly unobtrusive measures of actual behavior, can overcome biases of self-report	Behavioral records are often imprecise; imprecise translation between trace measures and psychological constructs or behavior

Ambulatory assessment	Rich time series data; physiological and spatial data that can complement self-reported daily diary measurement; does not rely on self-report	Equipment may break; costly to measure and track large numbers of individuals; sampling bias may result from differential willingness to participate
Population observation (e.g., “lost letter”)	Can draw a quantitative picture of a community as a whole; can be paired with experimental methods; highly unobtrusive	Impossible to connect data to individuals to test individual-level hypotheses
Observation of situation characteristics	Can provide taxonomy of situational variables that affect behavior	Relatively little theory on situation characteristics to guide the data collection

Experimental and Quasi-Experimental

Randomized experiment	Establishes causal relationships in a naturalistic setting that does not eliminate potentially important variables	Problems with “take up” of the randomly assigned treatment; treatment spillover and attrition occur more frequently in field compared to laboratory experiments
Encouragement design	Helps study interventions for which it is difficult to preserve a true control group (e.g., political movements)	Both encouragement and intervention must be effective to observe a relationship between the IV and the DV
Randomized	Useful for cases in	Spillover must be

rollout / waiting list design	which all of the population needs to receive treatment eventually; when control group is treated following a certain period of time, investigators can test whether the control group reacted to intervention in the same way as the treatment group	prevented using spatial or other types of treatment segregation; limited window for measurement because control eventually receives treatment
Downstream experiment	There exist many potential indirect effects of previous experiments; investigator does not need to administer the experiment personally	Bias resulting from differential rates of success at tracking down members of initial experiment; investigator cannot control content of the treatment or quality of implementation of experiment
Hybrid lab-field	Gains in control and precision by implementing treatment or measuring DVs in a more controlled, laboratory-like setting	The laboratory component is obtrusive
Regression discontinuity	Useful when randomization is not possible, but there is a clear decision point along a continuous measure of eligibility for treatment	The decision point regarding eligibility must be monotonically measured and rigid; the expectation of comparability between treatment and control is

	regarding who will receive the treatment and who will not	not as strong as the expectation of comparability between two groups formed by random assignment; less statistical power than random assignment
Interrupted time series analysis	Useful for assessing the impact of a treatment that affected an entire population	Treatment must occur at an observed point in time; dependent measurement must include true time series data, ideally data reaching back to the same date of the intervention one or two years earlier

Note: For additional designs, see Shadish, Cook, & Campbell (2002); Green & Gerber (2012).

Advantages and Disadvantages of Field Experiments

The advantages of field experiments build on the advantages of fieldwork more generally. They reveal causal relationships that hold up in the “rich thicket” of myriad social influences. Field experiments can also serve multiple research goals at once; working in the field on an experiment provides opportunities for qualitative observation that can inspire or refine future hypotheses. On a more personal level, running experiments in the field can be very rewarding thanks to the social interaction and engagement with practitioners with knowledge of and insight into the psychological constructs of interest to investigators.

Of course, it is important to keep in mind several disadvantages to field experimentation. Because behavior is constrained less in the field than in the laboratory, problems with participation, or “take up,” of the randomly assigned treatment, treatment spillover, and attrition occur more frequently. Certain types of interventions are more difficult to manipulate in field settings, or may be manipulated less precisely, such as cognitive or emotional states. In light of this fact, it is important to keep in mind that validity and precision are properties of

research programs as well as individual studies (Brewer & Crano, [Chapter 2](#) in this volume), and so field experiments can be profitably combined with other studies to answer questions that may be further out of reach in the field.

Finally, field experiments are more logistically challenging to launch and to manage compared to most laboratory experiments. Field experimenters must become bureaucrats, politicians, marketers, and public relations managers in order to organize, interact with, and appease all of the various people involved in the enterprise (or they must hire or collaborate with competent partners who can do so). Challenges include getting permission from institutional review boards (IRBs) and from participating organizations in the field and identifying the participant samples and means to reach out to them and measure their outcomes. In the following sections we provide some general practical tips for field research.

Practical Issues of Research in Field Settings

Permission from stakeholders in the field. The first practical hurdle to overcome when you have chosen a research site or a population is establishing a collaboration agreement with the relevant stakeholder. The stakeholder may be the administration of a park where you plan to observe people, the director of a prison where you plan interviews, the CEO of a company whose services you would like to observe or randomly assign, and so forth. In our experience, there are two cases for field research that you can present to these stakeholders. One concerns the value of contributing to scientific knowledge about their environment or enterprise. The second, and in our experience the far more persuasive case, concerns the research's potential to benefit the stakeholder. For example, it may be that a park administrator is interested to see a descriptive analysis of interactions in various park spaces, a prison warden wants your insights into social dynamics inside the prison, or company management wants to know whether your treatments can improve their sales or efficiency. Of course, you can only promise to share data that will not compromise the well-being or privacy of your participants.

It is fair, particularly when the scientific investigator represents a tax on the stakeholder's resources or time, to offer learning in return. This can include writing a nonacademic, brief summary of the study's findings. To alleviate anxieties that a study will be providing a “thumbs up / thumbs down” assessment of an organization's environment or services, it is also important to explain to the

stakeholder that psychologists are interested in processes as much as they are interested in outcomes. Thus, even if a treatment is found to have negative effects, your research can provide clues as to why that may be happening, thereby allowing the partner to address the problem productively. This point, and the point that it is ethical to test *whether* interventions are having a beneficial or harmful effect, is useful when partnering with organizations that seek to promote prosocial change in the world.

But even with all these points addressed, you may need to convey another type of assurance to stakeholders. They may need to feel confident that the researcher is supportive of their purposes. For instance, a while ago, one of us led a research team seeking to test certain theoretically relevant request strategies on blood donations. Arranging for the tests necessitated the cooperation of the local blood services organization and required that we convince their officers that a collaboration would be worthwhile not just to us but to their organization's vitality. Although we thought that we had made a compelling case in these regards, the organization's chief administrator hung back from authorizing our project. It was not until a junior member of his staff quietly informed us of the reason for her boss's reluctance that we understood what we had left out of our persuasive approach. "None of you has given blood yet," she whispered during a break in one of our meetings. Mildly chastised but properly enlightened, we asked just before the meeting's close how we might contribute to the organization's important goals by donating a pint or two of blood ourselves. An opportunity was arranged, blood was drained, and full approval of our project followed within the week.

Memorandum of understanding. When you have come to an agreement with a field-based stakeholder to conduct your research, it is advisable to draw up a simple memorandum of understanding regarding exactly what the study activities will entail, your ownership of the data, your intention to strip any participant identity from the data, and your right to publish the data. Sometimes it is wise to include a clause that you will omit the identity of your field site, if the stakeholder desires, and that you will share the data if the stakeholder has a use for it (with all identities protected as mandated by your IRB).

Institutional Review Boards. IRBs are sometimes much more reluctant to grant permission to conduct research in field settings, although this is quite varied from institution to institution. Our advice, particularly if you hope to conduct research in a setting that involves some degree of sensitivity, vulnerable populations, or danger, is to do your homework. Contact researchers who have

done work in similar settings and ask for a copy of their IRB application and approval. Propose similar safeguards in your own work and cite previous work that was approved for that setting. If you can, contact your IRB members ahead of time and ask if there are immediate issues that you should remember to address in your application. Potential issues include the possibility of obtaining fully informed consent, the question of how to recognize privacy rights in proposed observation, and the potential for embarrassing or compromising behavior to be recorded. Provide as much information as you can on the field setting, so that decisions are not made on the basis of too little information about the potential risks.

Conclusion

The prospect of adding field research to an existing program of laboratory research may trigger different reactions among social psychologists. Psychology students may worry about the reception of field research in their department or in journals where they hope to publish. Faculty may worry about the same issues, and additionally about the time or the learning curve involved in mounting a line of field research. To both groups of psychologists we emphasize once more that field research can be used for the *development* of psychological theory, not solely the (underappreciated) application of theory. Moreover, we expect theories developed with the aid of field research to be more psychologically and pragmatically consequential.

To students in particular, we add that the activity of choosing the right field setting and real-world variables will compel you to describe and make concrete your theoretical constructs of interest like no other requirement in your training program. To psychology faculty, we point to William McGuire's (1973) once more relevant advice: "if the budgetary cutbacks continue, instead of running ever faster on the Big-Science treadmill, we [should] ... rediscover the gratification of personally observing the phenomena ourselves and experiencing the relief of not having to administer our research empire" (p. 455).

Personally observing the psychological phenomena of interest may include creating laboratory simulations that are much more realistic. And certainly, all researchers should balance their research portfolios to include some fieldwork, some laboratory work, and some pure theoretical work. But we will end on a challenge: If we are correct that fieldwork creates the most accurate and consequential theoretical maps for psychology, then our field today finds itself in

a concerning state of imbalance. The volume of insights from laboratory work far outweighs those from the field. Conduct research in the field and create theoretical maps that will bring psychological science into its next era.

References

- Adams, G. (2005). The cultural grounding of personal relationship: Enemyship in North American and West African worlds. *Journal of Personality and Social Psychology*, 88, 948–968.
- Adams, G., & Stocks, E. L. (2008). A cultural analysis of the experiment and an experimental analysis of culture. *Social and Personality Psychology Compass*, 2, 1895–1912.
- Angrist, J., & Krueger, A. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15, 69–85.
- Aron, A., Melinat, E., Aron, E. N., Vallone, R., & Bator, R. (1997). The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and Social Psychology Bulletin*, 23, 363–377.
- Asch, S. E. (1952). *Social psychology*. New York: Prentice-Hall.
- Bator, R. J., & Cialdini, R. B. (2006). The nature of consistency motivation: Consistency, inconsistency, and anticonsistency in a dissonance paradigm. *Social Influence*, 1, 208–233.
- Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior?. *Perspectives on Psychological Science*, 2(4), 396–403.
- Beaman, L., Chattopadhyay, R., Duflo, E., Pande, R., & Topalova, P. (2009). Powerful women: Does exposure reduce bias? *Quarterly Journal of Economics*, 124, 1497–1540.
- Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2010). *Using mechanical Turk as a subject recruitment tool for experimental research*. Unpublished manuscript.
- Blackwell, L., Trzesniewski, K., & Dweck, C.S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, 78, 246–263.

- Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, 54, 579–616.
- Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, 24, 409–429.
- Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, 329, 1194–1197.
- Chatman, J. A., & Flynn, F. J. (2005). Full-cycle micro-organizational behavior research. *Organization Science*, 16, 434–447.
- Cialdini, R. B. (1980). Full-cycle social psychology. *Applied Social Psychology*, 1, 21–47.
- Cialdini, R. B. (1993). *Influence (rev): The Psychology of Persuasion*. New York: HarperCollins.
- Cialdini, R. B. (2003). Crafting normative messages to protect the environment. *Current Directions in Psychological Science*, 12(4), 105–109.
- Cialdini, R. B. (2009a). *Influence: Science and practice* (5th ed.). Boston: Allyn & Bacon.
- Cialdini, R. B. (2009b). We have to break up. *Perspectives on Psychological Science*, 4, 5–6.
- Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic indicators of psychological change after September 11, 2001. *Psychological Science*, 15, 687–693.
- Cook, S. (1978). Interpersonal and attitudinal outcomes in cooperating interracial groups. *Journal of Research and Development in Education*, 12, 97–113.
- Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.
- Crosby, F., Bromley, S., & Saxe, L. (1980). Recent unobtrusive studies of black and white discrimination and prejudice: A literature review. *Psychological Bulletin*, 87, 546–563.
- Ellsworth, P. C. (1977). From abstract ideas to concrete instances: Some guidelines for choosing natural research settings. *American Psychologist*, 32,

604–615.

- Emerson, R. M., Fretz, R. I., & Shaw, L. L. (1995). *Writing ethnographic fieldnotes*. Chicago: The University of Chicago Press.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson and Company.
- Fiske, S. T. (1992). Thinking is for doing: Portraits of social cognition from daguerreotype to laserphoto. *Journal of Personality and Social Psychology*, 63, 877–889.
- Gerber, A. S., Green, D. P., & Kaplan, E. H. (2004). The illusion of learning from observational research. In I. Shapiro, R. Smith, & T. Massoud (Eds.), *Problems and methods in the study of politics* (pp. 251–273). New York: Cambridge University Press.
- Glaser, J., Dixit, J., & Green, D. P. (2002). Studying hate crime with the internet: what makes racists advocate racial violence? *Journal of Social Issues*, 58(1), 177–193.
- Gneezy, A., Gneezy, U., Nelson, L., & Brow, A. (2010). Shared social responsibility: A field experiment in Pay-What-You-Want pricing and charitable giving. *Science*, 329, 325–327.
- Goel, S., Mason, W., & Watts, D. J. (2010). Real and perceived attitude agreement in social networks. *Personality and Social Psychology*, 99, 611–621.
- Goffman, E. (1959). *The presentation of self in everyday life*. New York: Anchor Books Doubleday.
- Gosling, S. D., & Johnson, J. A. (Eds.). (2010). *Advanced methods for conducting online behavioral research*. Washington, DC: American Psychological Association.
- Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust Web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, 59, 93–104.
- Green, D. P., & Gerber, A. S. (2002). Reclaiming the experimental tradition in political science. In I. Katznelson & H. Milner (Eds.), *Political science: The state of the discipline* (pp. 805–832). New York: Norton.

- Green, D., & Gerber, A. (2012). *Field experiments: Design, analysis, and interpretation*. New York: Norton.
- Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist*, 52, 1115–1124.
- Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect. *Journal of Personality and Social Psychology*, 82, 903–918.
- Henry, P. J. (2008). College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry*, 19, 49–71.
- Huggins, M. K., Haritos-Fatouros, M., & Zimbardo, P. G. (2002). *Violence workers: Police tortures and murderers reconstruct Brazilian atrocities*. Berkeley: University of California Press.
- James, W. (1981 [1907]). *Pragmatism* (B. Kuklick, Ed.). Indianapolis, IN: Hackett Publishing Company.
- Kelley, H. H., Holmes, J. G., Kerr, N. L., Reis, H. T., Rusbult, C. E., & Lange, P. A. M. (2003). *An atlas of interpersonal situations*. Cambridge: Cambridge University Press.
- Kenrick, D. T., Goldstein, N., & Braver, S. L. (Eds.). (2012). *Six degrees of social influence: Science, application and the psychology of Robert Cialdini*. New York: Oxford University Press.
- LaFrance, M., & Woodzicka, J. A. (2005). The effects of subtle sexual harassment on women's performance in a job interview. *Sex Roles*, 53, 67–77.
- Lareau, A., & Shultz, J. (Eds.). (1996). *Journeys through ethnography: Realistic accounts of fieldwork*. Boulder, CO: Westview Press.
- Lewin, K. (1944/1997). Problems of research in social psychology. In *Resolving social conflicts & field theory in social science*. Washington, DC: American Psychological Association.
- Lewin, K. (1947). Frontiers in group dynamics: concept, method and reality in social science; social equilibria and social change. *Human relations*.
- Linder, D. E., Reich, J. W., & Braver, S. L. (2012). Collective full cycle social psychology: Models, principles, experience. In D. T. Kenrick, N. Goldstein, &

- S. L. Braver (Eds.), *Six degrees of social influence: Science, application and the psychology of Robert Cialdini* (pp. 213–230). New York: Oxford University Press.
- McGuire, W. J. (1969). Theory-oriented research in natural settings: The best of both worlds for social psychology. In M. Sherif & C. W. Sherif (Eds.), *Interdisciplinary relationships in the social sciences* (pp. 21–51). Chicago: Aldine Publishing Company.
- McGuire, W. J. (1973). The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology*, 26, 446–456.
- Mortensen, C. R., & Cialdini, R. B. (2010). Full-cycle social psychology for theory and application. *Social and Personality Compass*, 4, 53–63.
- Paluck, E. L. (2009). Reducing intergroup prejudice and conflict using the media: A field experiment in Rwanda. *Journal of Personality and Social Psychology*, 96, 574–587.
- Paluck, E. L. (2010a). Is it better not to talk? Group polarization, extended contact, and perspective-taking in eastern Democratic Republic of Congo. *Personality and Social Psychology Bulletin*, 36, 1170–1185.
- Paluck, E. L. (2010b). The promising integration of field experimentation and qualitative methods. *The ANNALS of the American Academy of Political and Social Science*, 628, 59–71.
- Paluck, E. L., Blair, G., & Vexler, D. (2011). *Entertaining, informing, and discussing: Behavioral effects of a democracy-building radio intervention in Southern Sudan*. Working paper, Princeton University.
- Pennebaker, J. W., Mehl, M. R., Crow, M. D., Dabbs, J., & Price, J. H. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, 33, 517–523.
- Rathje, W. L., & Hughes, W. W. (1975). The garbage project as a nonreactive approach: Garbage in...garbage out? In H. W. Sinaiko & L. A. Broedling (Eds.), *Perspectives on attitude assessment: Surveys and their alternatives* (pp. 151–167). Washington, DC: Smithsonian Institution.
- Reich, J. W. (2008). Integrating science and practice: Adopting the Pasteurian Model. *Review of General Psychology*, 12, 365–377.

- Reis, H. T. (2008). Reinvigorating the concept of situation in social psychology. *Personality and Social Psychology Review*, 12, 311–329.
- Reis, H. T., & Gosling, S. D. (2010). Social psychological methods outside the laboratory. In S. T. Fiske, D. T. Gilbert & G. Lindzey (Eds.), *Handbook of social psychology: Volume 1* (5th ed., pp. 82–114). Hoboken, NJ: John Wiley & Sons.
- Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, 5, 2–14.
- Rubin, D. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100, 322–331.
- Salganik, M. J., & Watts, D. J. (2008). Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social Psychology Quarterly*, 71, 338–355.
- Sears, D. O. (1986). College sophomores in the lab: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51, 515–530.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized casual inference*. Boston, MA: Houghton Mifflin Company.
- Shadish, W., Galindo, R., Wong, V., Steiner, P., & Cook, T. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods*, 16, 179–191.
- Shulman, A. D., & Berman, H. J. (1975). Role expectations about subjects and experimenters in psychological research. *Journal of Personality and Social Psychology*, 32, 368–380.
- Sondheimer, R. M., & Green, D. P. (2010). Using experiments to estimate the effects of education on voter turnout. *American Journal of Political Science*, 54, 174–189.
- Tomlinson, M., Rotheram-Borus, M., Doherty, T., Swendeman, D., Tsai, A., le Roux, I., Jackson, D., Chopra, M., Stewart, J., Ijumba, P. (2011) Mobile technologies train, monitor, and support community health workers for improved quality of care. Working paper.

- Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology*, 92, 82–96.
- Webb, E., Campbell, D., Schwartz, R. & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. B. (1981). *Nonreactive measures in the social sciences* (2nd ed.). Boston: Houghton Mifflin Company.
- West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Casual inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–84). Cambridge: Cambridge University Press.
- West, S. G., & Graziano, W. G. (2012). Basic, applied, and full cycle social psychology: Enhancing casual generalization and impact. In D. T. Kenrick, N. Goldstein, & S. L. Braver (Eds.), *Six degrees of social influence: Science, Application, and the psychology of Robert Cialdini* (pp. 181–202). New York: Oxford University Press.
- West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on casual inference. *Psychological Methods*, 15(1), 18–37.
- Wilson, T. D., Aronson, E., & Carlsmith, K. (2010). The art of laboratory experimentation. In S. T. Fiske, T. D. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology: Volume 1* (5th ed., pp. 51–81). Hoboken, NJ: John Wiley & Sons.

¹ Although a full-cycle approach to social psychology is not an exclusively field-based orientation, field investigation is a prominent and important component. The advocated process typically involves (1) recognizing a powerful and interesting phenomenon in the natural environment, (2) conducting an initial scientific test of the validity of the phenomenon, usually in the field, (3) conducting further scientific investigation of the mediating processes and theoretical underpinnings of the effect, often in the lab, and (4) looking back to naturally occurring situations to assess the match between the characteristics of

the effect as it appeared in the studies and how it appears in the real world. As an upshot, one can better determine the presence, reliability, and force of psychological phenomena in the real world while identifying the psychological mechanisms underlying these phenomena. Naturalistic observation, field research, and laboratory research become symbiotic, with strengths and weaknesses that complement one another (Cialdini, [1980](#); Mortensen & Cialdini, [2010](#)).

Part two Procedural Possibilities

Chapter six Using Physiological Indexes in Social Psychological Research

Jim Blascovich*

Social psychologists, like other behavioral scientists, have long been drawn to objective measures of psychological constructs. The value of these measures largely derives from avoidance of problems stemming from subjective self-report and human observational measures, such as participant and observer biases. That objective measures can often be collected, recorded, and scored via technology-based data collection systems is also appealing.

One subcategory of objective measures, neurophysiologically based indexes of social psychological constructs, has risen in popularity over the past three decades. During that time, the rationale for the use of physiological measures was transformed from one based largely on a naïve mystique¹ driven by overly simplistic notions of the bodily component of mind-body relationships to one based on a more sophisticated understanding of their complexity. That is, social psychologists have developed a more sophisticated understanding of human neurophysiology supporting particular configurations of covert, largely automatic and/or uncontrollable bodily processes as indexes of social psychological constructs including, especially, affective, cognitive, and motivational ones.

Today, the terms “social psychophysiology” and the more recent “social neuroscience” refer to indexes based on human biological systems controlled by the peripheral and central nervous systems, respectively. This chapter focuses on the former while Berkman, Cunningham, and Lieberman's contribution (Chapter 7) to this volume focuses on the latter.

Social psychologists often successfully test hypotheses using validated neurophysiological response patterns associated with important psychological constructs and processes instead of, or in addition to, more traditional self-report and behavioral indexes. The methodological tools to distinguish physiologically among superordinate psychological states enable researchers to critically evaluate models that assume the existence of one or more such states as a

function of theoretically specified circumstances and provide more power to multimethod triangulation involving self-report and observational methods. This chapter focuses on successful exemplars of such indexes.

Today, social psychophysiological indexes allow researchers to objectively and accurately distinguish, for example, appetitive from aversive motivational states, positive from negative affect, attention from inattention, and linguistic from nonlinguistic processing. Furthermore, these indexes are based on participant responses generally outside of their conscious control. Because these indexes are typically continuous, fluctuating coterminously with underlying dynamic psychological states over time, they provide an important temporal perspective to data and, in turn, to theory. Such indexes allow social psychologists to test the motivational, affective, and cognitive underpinnings of a wide range of theories within typical empirical contexts guided by powerful experimental designs (cf. Blascovich & Mendes, 2010). The objective in this chapter is to help readers understand the nature of neurophysiological processes and the utility of physiological measures as state-of-the-art empirical indexes of constructs fundamental to social psychological theories. Casual readers should benefit as consumers of social psychological research that includes such measures. More thorough and resourceful readers should be able to implement physiological measures in their own research.

In this chapter, we cover relevant background information, including the evolution of social psychophysiology, a brief discussion of relevant epistemological issues, and the nature of physiological indexing. We also briefly review and integrate general information regarding physiological control processes and general technological approaches to their measurement. We move on to a brief discussion of threats to validity in physiological measurement. Next, we present illustrations of state-of-the-art physiological indexes of important motivational and affective constructs. Finally, we conclude with a general assessment and summary.²

Background Information

The Evolution of Social Psychophysiology

History.

Although the use of heart rate to index interpersonal attraction can be traced to

ancient Greek and Roman physicians, social psychophysiological forays appeared early in the 20th century. Researchers such as Lasswell (1936), Mittleman and Wolff (1939), and Boyd and DeMascio (1954) explored the relationships of specific psychological constructs, such as motivation, speech rate, emotions, and the nature of interpersonal relationships to specific unitary physiological responses including respiration, pulse rate, finger temperature, and skin conductance. These early explorations, though mostly eventual failures substantively, exemplified the advanced thinking of at least some investigators who believed that combining biological and psychological approaches could provide useful information to proponents of each.

However, these failures provided valuable information if only by highlighting the shortcomings of their epistemologies. Hence, we can learn from the early history of social psychophysiology by investigating its failure to fulfill the hopes of its early adherents. They failed for many reasons: naive integration of relatively unsophisticated physiological and psychological theoretical frameworks, weak logical and epistemological bases for drawing inferences, and primitive methodologies and technologies. Yet, in some sense, these weaknesses, errors in a sort of transcendent methodological trial-and-error scheme, provided necessary steps in the evolution of social psychophysiology. Consequently, these weaknesses represent problems that we can learn about and avoid.

The promise of physiological indexes drove remedies for problems inherent in the integration of social psychology and psychophysiology. These began appearing in the literature with Shapiro and Crider's (1969) classic *Handbook of Social Psychology (2nd Edition)* chapter on physiological approaches to social psychology. Cacioppo and Petty's (1983) and Waid's (1984) edited volumes, respectively entitled *Social Psychophysiology* and *Sociophysiology*, showcased the works of newer, more fruitful investigations integrating social psychological and psychophysiological methodologies. The intensive summer "Program for Advanced Study and Research in Social Psychophysiology," sponsored by the National Science Foundation and led by John Cacioppo and his colleagues (summers of 1986 through 1990), provided more than 60 social psychological researchers and those in related disciplines a firm grounding in psychophysiological theory and measurement techniques as well as in important logic, epistemological, and physiological data analytic techniques. More recently, the fifth edition of the *Handbook of Social Psychology*, after a three-edition hiatus, reprised social psychophysiology (Blascovich & Mendes, 2010) as a topic deserving a chapter.

Function.

Social psychophysiology is regarded in the broadest sense as a method. Its value for social psychological researchers lies primarily in its provision of a growing set of objective measures or indexes of theoretical and empirical constructs. To be sure, valid use of these indexes requires important background information about physiological theory and its attendant assessment technologies. Fortunately, more and more sophisticated information of this type has become available (cf. Blascovich, Vanman, Mendes, & Dickerson, 2011).

The advantages of social psychophysiological method accrue from its stipulation of relatively unbiased, real-time indexes of psychological processes related to motivational, affective, and cognitive constructs – ones that are often difficult to quantify without bias or artifact in the absence of such objective assessments and impossible to quantify, especially temporally, otherwise in vivo. The disadvantages accrue from difficulty in attaining appropriate background information, specialized training, and relative cost. Although perfectly adequate and less cumbersome equipment continues to become more reasonably priced, utilizing social psychophysiological indexes usually involves an investment in equipment above that involved in utilization of more traditional social psychological methodologies. However, as advances in instrumentation have played a key and heuristic role in the history of all sciences, failure to secure and use it can render one's work obsolete.

Epistemological Issues

The gradual rejection of Cartesian dualism by many life scientists opened up new avenues of exploration linking mind and body (cf. Damasio, 1994). The emergence and growth of new and intellectually exciting fields such as social neurophysiology, psychoneuroimmunology, and psychoneuroendocrinology give testimony to the value of multilevel or multisystemic approaches for understanding the interconnected nature of body and mind. Social psychophysiology, like these other approaches, assumes the *identity thesis*. Specifically, social psychophysiolgists assume that biological structures and physiological processes embody all human behaviors, including thoughts and feelings as well as overt actions. Consequently, researchers can turn to these structures and processes in order to learn more about social behavior (Cacioppo & Tassinary, 1990b).

Unfortunately, a one-to-one correspondence between specific behaviors and

unitary physiological responses rarely exists (one-to-one relationships rarely exist for self-report or behavioral indexes either). This lack of singular correspondence derives both from the multifunctionality of physiological processes and the complexity of behavior, especially social behavior. To take a simple example, heart rate generally increases during overt physical activities such as aerobic exercise, but it also increases during covert mental activities in the absence of metabolic demand such as anticipation of the arrival of a romantic partner at an airport, completion of a written examination, or speech preparation. Thus, an increase in heart rate, or any other “single” physiological response for that matter, typically fails to unambiguously index specific behaviors across contexts.

In order for biological responses to be useful methodologically to social psychologists, both the biological and psychological contexts within which these responses occur must be understood. Unfortunately, sufficient understanding and familiarity with the biological context has eluded many social psychologists in the past. Historically, social psychologists, like many others, simply assumed the interchangeability of autonomic and somatic physiological responses such as heart rate, blood pressure, skin conductance, and muscle tension as indexes of emotion. Fortunately, we see very little of this rather naive approach reported in the social psychological literature today.

In the past, psychophysiolgists – and today even some neuroscientists (as opposed to social neuroscientists) – pushed aside much of the social psychological context as error. Specifically, psychophysiolgists traditionally assumed that error variance in the relationships between specific behaviors and specific physiological responses derive not only from random measurement error but also from the contributions of systematic individual and situational influences. They label the former as individual response stereotypy (i.e., typical of the specific individual) and the latter as situational response stereotypy (i.e., typical of the specific situation), thereby lumping all individual differences in responses to similar situations and differences in responses between situations as sources of error. Psychophysiolgists do not assume the interchangeability of psychologically related physiological responses, and social psychologists do not treat individual and situational influences on any kind of responses as uninteresting or as an error. Social psychophysiolgists should do neither. The researcher employing a social psychophysiological approach must base his or her assumption of the critical identity thesis on thorough knowledge of its biopsychosocial underpinnings.

The Nature of Physiological Indexes of Psychological Constructs

Invariance defines the ideal relationship between a construct and its index. At the nominal level of measurement, invariance means that the construct and the index always co-occur. If the construct is there, so is the index and vice versa. For example, immunologists often index the occurrence of specific viral infections by the presence of specific viral antibodies. If the individual shows any evidence of the antibody, immunologists assume that infection has occurred. If the individual shows no evidence of the antibody, they assume the individual has never been infected. At the ordinal level of measurement, invariance means that the construct and the index always co-occur and covary in a ranked or ordinal manner. For example, angiographers traditionally indexed coronary artery disease using ordinally increasing categories of occlusion for each coronary artery that they assessed from cineangiographic (or “moving”) X-rays of the coronary arteries in vivo. Consequently, level 3 occlusion indexes more disease than level 2 occlusion and less than level 4 occlusion. At the interval-ratio level, invariance means that the construct and the index always co-occur and covary monotonically. For example, exercise physiologists index muscle movement from integrated muscle action potentials. The greater the integrated muscle action potential, the greater the muscle flexion.

Physiological invariants of psychological constructs have proven difficult to establish for at least two reasons. First, invariant indexes, whether subjective or objective, of social psychological constructs often prove elusive because the target constructs – for example, risk-taking, love, prejudice, or the self-concept – themselves prove difficult to define (Blascovich & Ginsburg, 1978). Second, a one-to-one correspondence between specific psychological constructs and unitary physiological responses rarely exists. Both domains are complex. Nevertheless, one can still devise valid physiological indexes or indicators of psychological constructs, perhaps even invariant ones.

We contend, in general, that as one narrows or limits the behavioral construct and context while expanding the constellation of relevant physiological responses forming the index, one can approach the one-to-one correspondence or invariance necessary for the development of valid physiological indexes of the social psychological constructs. Expanding the constellation of physiological responses can be accomplished by examining multiple physiological responses over time (Blascovich & Kelsey, 1990; Cacioppo & Tassinary, 1990a).

Consequently, a constellation of physiological responses can constitute an index that serves as a marker, if not an invariant one, of context-specific psychological constructs.

Building on the work of Troland (1929), Cacioppo and colleagues (e.g., Cacioppo, Tassinary, & Berntson, 2007) described the general nature of relationships between specific behavioral constructs and specific physiological responses in five categories: one-to-one, one-to-many, many-to-one, many-to-many, and null (see Figure 6.1). As implied earlier, one-to-one relationships form the basis for meaningful and specific physiological (as well as other types) indexes of psychological constructs. In a sense, a one-to-one relationship represents the goal for development of social psychophysiological indexes.

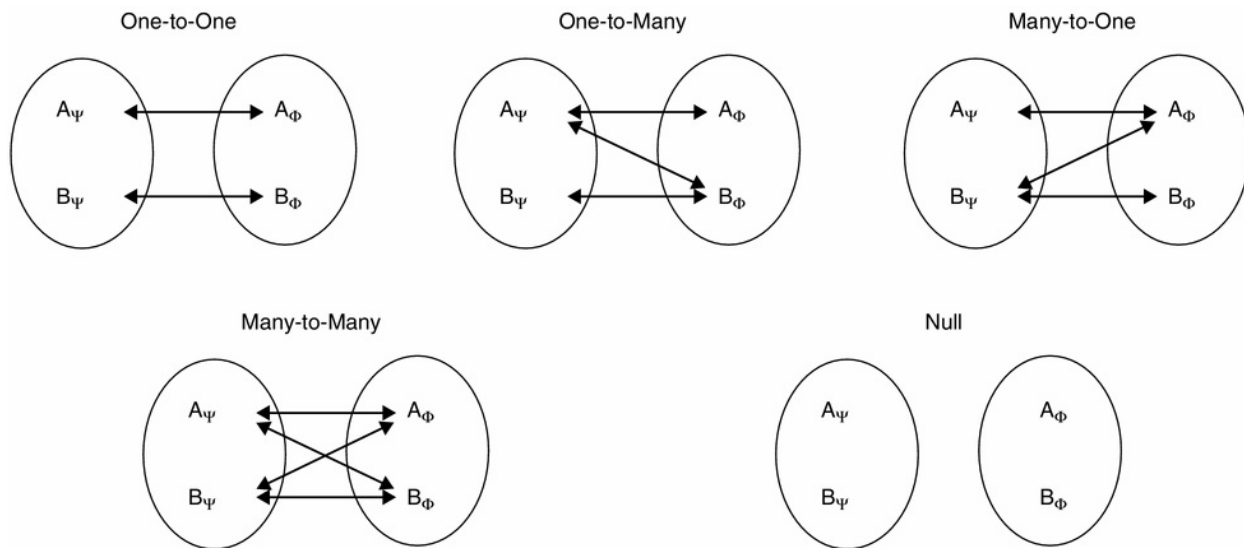


Figure 6.1. General relationships between psychological constructs (ψ) and physiological responses (ϕ).

Figure 6.1 (upper left panel) depicts one-to-one relationships between different psychological constructs (represented by A_ψ and B_ψ) and different sets of physiological responses (represented by A_ϕ and B_ϕ) symbolically. The development of valid and useful physiological indexes of the constructs is more likely if the following four propositions hold:

1. *The psychological constructs, A_ψ and B_ψ are conceptually distinct.* This proposition assumes the necessity of appropriate conceptual analysis and operational definition of the constructs of interest. One cannot index a psychological construct until one defines it explicitly and operationalizes it validly (Blascovich & Ginsburg, 1978). To the extent

that psychological constructs are conceptually clear and nonoverlapping (e.g., threat and challenge; joy and sadness) rather than overlapping (e.g., threat and fear; compassionate and companionate love), distinctive physiological indexes are more likely. This proposition applies to physiological and nonphysiological indexes alike. For this reason, physiological indexes are as unlikely as any to allow us to differentiate easily related but fuzzy concepts such as liking and loving, sadness and depression, peripheral and heuristic cognitive processing, achievement and intrinsic motivation, and so forth. Furthermore, the strategy of equating a psychological construct with a specific set of physiological responses creates the problem of definitional operationism (e.g., intelligence is what intelligence tests measure) unless an invariant relationship has been demonstrated.

2. *The sets of physiological responses, A_ϕ and B_ϕ , are each more inclusive (e.g., cardiac, hemodynamic, and vascular) rather than less inclusive (e.g., heart rate).* According to this proposition, although a single physiological response comprises the logically minimal response set defining a physiological index, given the many interrelationships among psychological and physiological processes, sets including two or more responses are considerably more desirable, if not essential. Two major arguments support this proposition. First, statistically, n responses occurring in predictable ways are less likely to occur by chance than $n-1$ responses. Hence, the problem of Type I error becomes reduced, and the basis for inference becomes stronger. Second, because the body's physiological systems predictably work in concert, sometimes in a complementary and sometimes in an oppositional fashion, the greater the number of physiological responses comprising an index, the more convergent and divergent validation of the psychological construct they provide.
3. *The sets of physiological responses, A_ϕ and B_ϕ , overlap in substance but not in form.* Stronger inference results when differing internal patterns of a common set of physiological responses, rather than different sets of physiological responses, distinguish the psychological constructs. Hence, differential patterns involving the same sets of physiological responses provide distinguishing information. From a purely logical point of view it matters little whether every physiological response in a set differs from its counterpart in the other set, only that at least one does.

4. The sets of physiological responses, A_φ and B_φ , are assessed continuously over time. Continuous time-series assessment of physiological responses increases the likelihood of distinguishing differential patterns of responses between sets of overlapping responses above and beyond that of single time-point samples of each response or even averaged n -point samples. For example, multiple time-point assessments allow us to discern patterns involving linear increases versus decreases, accelerations versus decelerations, and polynomial trends versus linear ones, whereas, single time-point samples or averaged n -point samples do not. Figure 6.2 illustrates the incorporation of these four propositions into the graphic illustrating one-to-one relationships.

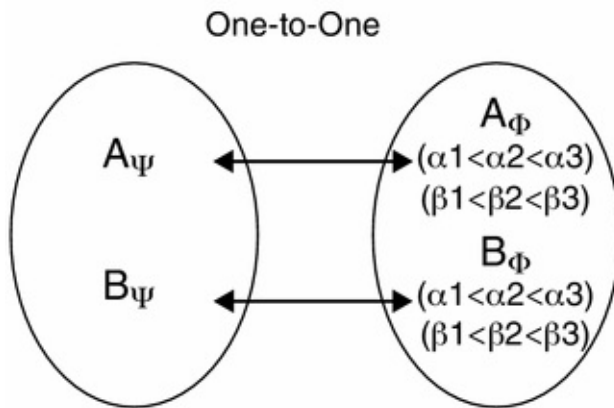


Figure 6.2. Expanded notion of one-to-one relationships between constructs (ψ) and physiological responses (φ).

Logically, it is not a necessary condition of valid indexing that any of these propositions hold. Single-response indexes are logically possible and arguably valid, as described later in the chapter. Second, conditions satisfying all four propositions are not sufficient for valid indexing. Thus, even if all these propositions hold, a reasonable basis for one-to-one relationships and, hence, indexes is not achieved unless a reasonable theoretical basis for relating the specific construct with a specific pattern of physiological responses exists. However, satisfaction of these propositions increases the confidence with which we can theoretically propose or empirically apply an index.

As we illustrate later in the chapter, the development of an appropriate theoretical basis for a physiological index may occur deductively, inductively, or as a combination of the two. That is, one may derive or deduce differential patterns among physiological responses as a function of the mediation of

distinctive neuropsychological processes on the basis of existing physiological and social psychological theory, or one may establish the association of specific patterns with specific psychological processes and develop and test a theoretical explanation or basis for the relationship.

Basic Physiological Processes

Understanding bodily processes and their relevance to psychological processes is advancing at a relatively rapid pace. Likewise, new technologies appropriate to observation and recording of bodily processes appear constantly. The more we know about each, the better we can justify, develop, and implement psychophysiological indexes of psychological constructs and processes. Clearly, those interested can avail themselves of more sophisticated physiological indexes of psychological constructs today than a decade or two ago. Just as clearly, those interested will be able to use even more sophisticated indexes in the future. In this section we can only sketch some of the important aspects of physiological processes. We encourage readers who want to use physiological indexes to immerse themselves in physiological background material from excellent sources such as Cacioppo, Tassinary, and Berntson (2007) and specialty sources as necessary.

The ways in which the body meets the requirements of life maintenance and environmental demands continue to amaze. Biochemical, electrochemical, hydraulic, and mechanical processes operate in fantastically complicated but well-organized and integrated ways to operate and maintain anatomical structures as well as life-sustaining (e.g., metabolic) processes efficiently and to produce behaviors involving simultaneous psychological processes, including motivation, cognition, affect, and movement.

The body can be described from a cybernetic systems or subsystems perspective dichotomizing bodily systems, albeit fuzzily, into control and operational ones. The former include neural, endocrine, and immunological systems. The latter include, for example, cardiovascular, digestive, electrodermal, respiratory, and somatic systems. Both system types involve basic cellular and intercellular tissues and processes. Hierarchically, the neural systems generally control or at least influence the other control systems directly and the operational systems both directly and indirectly through the endocrine and immunological systems, although endocrine and immunological control systems also may influence neural systems directly.

Control Systems

Neural Processes

Structure and Function.

The structure of the neural system has traditionally been organized anatomically because gross neural functions generally follow structure and location. However, readers should note that even the gross components of the neural system are quite well integrated and that various neural control mechanisms are not necessarily specific to different anatomical neural structures. Thus, the classic structures and substructures of the nervous system, such as the central nervous system (brain and spinal cord), autonomic nervous system (sympathetic and parasympathetic), and somatic nervous system, do not operate autonomously.

The nervous system functions in large measure as a communication and control system. Nearly every part of the body communicates with the brain by sending signals to the central nervous system via afferent peripheral nerves. Peripheral nerves enter (or “project” to) the brain via the spinal cord and brain stem. The brain internally transfers these signals to reception areas called somatosensory cortices. Projections from the somatosensory cortices directly and indirectly signal (i.e., communicate with) other areas and structures of the brain that “interpret” these signals and, in turn, generate outgoing neural (i.e., efferent) and endocrine signals that control distal or peripheral physiological processes. The control areas of the brain are somewhat specialized and include areas such as the amygdala and hypothalamus, involved in control of visceral and other autonomically controlled organs and areas, and the motor cortices and subcortical motor nuclei, involved in control of the musculoskeletal system.

Cellular Processes

Single-cell neurons provide the basic building blocks of the nervous system. Hundreds of billions of neurons exist overall, with the vast majority located in the brain. Configurations of neurons allow signal transmissions within and between the peripheral and central nervous systems, transmissions fundamental to macro-level neural processes. Although neurons can be distinguished on the basis of several dimensions, including size, length, location (e.g., central or peripheral), and synapses (i.e., connections), all neurons operate in much the

same general way, receiving and transmitting biochemical and bioelectric signals organized and generated by bodily or brain structures or by other neurons. Neuronal cell structure and physiology facilitates this function.

Endocrine Processes

The relatively simple organization and structure of the endocrine system belies its power. Upon neural stimulation, the pituitary gland generally initiates endocrine processes vis-à-vis the release of specific target chemical substances known as hormones into the blood stream, which, in turn, stimulate various bodily tissues including neural and other endocrine tissues. The various specific endocrine glands secrete still other hormones and neurotransmitters that affect various physiological processes. Direct neural stimulation of specific endocrine glands also occurs. The endocrine system functions directly to regulate growth and maturation and indirectly to modulate neural control of various operational systems. Psychophysicologists and others recognize the latter function as quite important, and its importance to social psychologists has expanded (Dickerson & Kemeny, 2004) .

Technological Background

Application of the sensors connecting participants' bodies and electronic physiological recording equipment requires technological expertise. However, such sophistication lies within the grasp of social and personality psychologists, and its value far outweighs its cost.

Here, we focus on an overview of “dry” technology, namely the technology of electrophysiological recording. This focus does not devalue the worth of “wet” technology, namely the technology of biochemical analysis within the endocrine-based social psychophysiological approach. Rather, it merely reflects the most popular technology employed to date by social psychologists using psychophysiological indexes and the technology most appropriate to the indexes of motivation and affect described later (but see Blascovich et al., 2011 for a discussion of endocrine methodology).

The technology of electrophysiological recording includes the steps between acquisition of physiological response signals and their recording (i.e., the “signal path”). As depicted in Figure 6.3, in the most complete case, physiological responses or signals are sensed, transduced, conditioned, and recorded. Successful implementation of physiological indexing of psychological constructs

requires a conceptual understanding on the part of the investigator of what occurs technologically at each step in the signal path. Fortunately, less detailed technical operating knowledge is required as the availability, sophistication, and user-friendliness (i.e., automation) of physiological recording equipment is continuously improving.

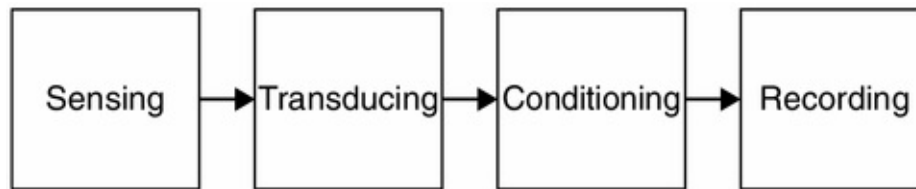


Figure 6.3. Signal path.

Physiological Response Signals

Historically, psychophysicists have focused their technological expertise primarily on noninvasive measurement techniques, which are ways of recording internal physiological responses of interest from the surface of the body. The success of this approach stems in no small part from the fact that various physiological events or responses produce detectable signals on the skin surface of the body, including both electrical and nonelectrical ones. The former include actual changes in electrical potentials as a function of specific physiological processes (e.g., changes in electrical potential across the heart as it completes its cycle, changes in muscle action potentials as muscle bundles contract). The latter include changes in pressure (e.g., intra-arterial blood pressure, intraocular pressure), movement (e.g., heart, lungs, digestive organs, blood flow), temperature, and tissue (e.g., production of sweat).

Signal Path

The path of physiological response signals for measurement purposes leads from the surface of the body to recording (see [Figure 6.3](#)). The signal path for electrical signals requires three steps, whereas the signal path for nonelectrical signals requires four steps. The additional step, signal transducing, changes nonelectrical physiological signals (such as respiration-caused expansions and contractions of the chest) into analog electrical signals.

Sensors

Electrodes are used as sensors for electrical physiological signals. Although varied in size, shape, and electrical conducting characteristics according to their

intended purpose (i.e., specific signal of interest), all electrodes function identically in terms of picking up changes in electrical potentials. If some part of the body (e.g., the heart, a muscle) emits changes in electrical potential of interest to the investigator, careful placement of electrodes will allow the investigator to optimize recording of that electrical potential.

Sensors for nonelectrical physiological signals include electrodes and other devices. Electrodes in this case help researchers discern tissue changes not from changes in electrical potentials emitted by bodily organs, but rather by enabling the investigator to deliver safe levels of electrical current to tissues, which by nature of specified physiological changes (e.g., increased levels of sweat in a sweat duct, blood flow through the heart) reliably influence current flow and measure changes in such flow. Using similar principles, other sensing devices help researchers discern physiological changes by delivering nonelectrical stimulation to the body and sensing changes in physiological response to such stimulation. For example, a photoplethysmograph uses a small lamp to deliver light to the surface of the skin (e.g., on the distal phalange of a finger or on the earlobe) combined with a photosensitive cell a short distance away sensing changes in light diffusion through the skin as a function of blood flow. Various other devices sense important nonelectrical physiological signals, including movement-sensitive devices such as strain gauges and temperature-sensitive devices such as thermistors.

Transducers.

As mentioned earlier, nonelectrical physiological response signals (e.g., blood flow, sweat levels) must be converted or “transduced” to analog electrical (i.e., voltage) signals. This occurs early in the signal path. Most transducers operate according to principles of electrical bridging. Bridge circuits produce a continuous voltage signal representing bioelectrical physiological responses measured with electrode sensors such as skin conductance and thoracic impedance. If one uses a fluctuating physical resistance device driven by an apparatus sensitive to nonelectrical physiological responses (e.g., strain gauge, thermistor, photoelectric cell), a bridge circuit will similarly produce a voltage analog signal corresponding to the underlying physiological cause of the nonelectrical physiological responses (e.g., movement, temperature, blood flow).

Conditioning

The next step in the signal path alleviates signal acquisition problems stemming from two factors: signal specificity and strength. The former relates to other physiological potentials and to ambient electronic noise. Here the researcher must be able to focus on the signal of interest despite the fact that the myriad organs and tissues of the body constantly produce electrical signals that become diffused throughout the body as they near its surface, and despite the pervasiveness of ambient electronic noise generated by electrical equipment. Generally, the researcher must also magnify the signals of interest in order to record them because recording devices generally require more powerful input signals than the body produces. Physiological potentials range from microvolt (0.000001 V) to millivolt (0.001 V) levels depending on the target physiological response. Many nonelectrical signals also require amplification because voltage and current levels applied to bodily tissue to assess changes in physiological responses such as sweat or blood volume activity must necessarily be quite weak for safety purposes.

Signal filtering and amplification represent the two primary modes of dealing with the problems of signal specificity and signal strength. Because physiological response potentials of interest to psychologists generally cycle at different frequencies, ranging from as much as 500 Hz (cycles per second) for muscle action potentials to less than 1 Hz for cardiac potentials and less than 0.1 Hz for gastric contractions, psychophysicists use electronic filters to prevent signals outside the frequency range of the target signal from obfuscating the signal of interest. Electronic amplification provides psychophysicists with the tools necessary to boost the strength of signals without altering signal topography.

Recording

The last step in the signal path requires storage. Historically, various devices have performed this function. Indeed, the polygraph takes its name from the first of these, the multichannel paper recording device. This device transformed analog voltage signals into pen excursions, which recorded voltage changes as waveforms on graph paper moving right to left under the pens. Actual manual measurements of various aspects of these waveforms provided necessary data values. With the advancement of electronic technology, analog voltage waveforms could be stored on magnetic tape. Today, widely available laboratory computers incorporating analog-to-digital converters allow online and high-fidelity digitization of analog waveforms, allowing recording on digital mass

media storage devices. Software algorithms have replaced manual measurement, thereby at least partially automating scoring of these analog signals and decreasing data errors.

Specific Methodological Concerns

In addition to knowledge of specialized signal path technology, proper use of psychophysiological indexes in experiments requires attention to specific threats to validity. Although social psychologists have long known well the labels and the logical and substantive bases for many such threats (thanks in large part to Campbell & Stanley, 1963), we generally lack familiarity with their manifestations within the psychophysiological domain. In addition to general concern regarding all the threats to validity, researchers must take special care to ameliorate the effects of maturation, testing, and instrumentation when using psychophysiological indexes.

Threats to Validity

Maturation.

Maturation poses special considerations for psychophysiological researchers not only in terms of gross cross-sectional maturational differences among research participants but also in terms of relatively short-lived within-and between-participant differences. Regarding the former, one should usually avoid including participants from vastly different age groups (e.g., adolescents, senior citizens) in the same study because the nature of many psychophysiological meaningful responses often changes dramatically over the life span, and such heterogeneity of research participants would increase the within-condition variability of physiological recordings. For example, muscles atrophy with age. Using certain somatic electromyographic (EMG) measures of teenagers and octogenarians could increase the ratio of unwanted to meaningful variance to levels rendering the statistical power of a design quite low.

Perhaps less obviously, biological processes related to fatigue, digestion, drug intake, environment, and so on moderate physiological and therefore psychophysiological responses. Investigators employing psychophysiological indexes should take into account normal individual and diurnal maturational variations in various behaviors such as eating, drinking, exercise, and sleep, as well as environmental factors including temperature and humidity.

Experimentally controlling individual and diurnal variations in metabolic processes clearly requires not only random assignment of participants to experimental sessions during normal, nonprandial waking hours but also often requires specific instructions regarding sleeping, eating, and drug-taking (e.g., alcohol, caffeine, medications, nicotine) behaviors prior to experimental sessions.

Testing.

Human physiological systems usually adapt to environmental stimuli and demands. Novel stimuli and demands generally elicit stronger physiological responses than familiar ones do. Physiological responses typically habituate to repeated stimulus presentation or situational demands relatively quickly, often within a couple of minutes. Consequently, if repeated stimulus presentations or behaviors (including cognitions) prove experimentally necessary, one must often control for habituation effects. In social psychological investigations, this can usually be accomplished by changing the content or form of the stimuli or task demands. For example, cardiovascular responses to a serial subtraction task will quickly return to baseline unless the subtractant is periodically changed.

Another unique effect of testing in psychophysiological experiments involves iatrogenic changes in the nature of the target variable or measure itself. For example, repeated blood pressure readings involving occlusive blood pressure cuffs can temporarily compress vascular tissue underneath the cuff, causing less than normal vascular elasticity, which in turn leads to erroneous blood pressure readings as well as the discomfort of experimental participants. The solution involves either decreasing the frequency of such intrusive measurement procedures, allowing underlying tissues to recover fully between measurements, or the use of minimally intrusive monitoring equipment (which in most cases, including blood pressure equipment, is available).

Instrumentation.

Although physiological measurement apparatuses have become substantially more reliable and less susceptible to environmental “noise” than even a few years ago, operating characteristics of such equipment as well as those of the measurement environment can and do change. A controlled, noise-free (electronically and otherwise) environment optimizes physiological recording reliability. A well-trained, noise-free operator who understands the nature of the measurement devices, proper participant hookup, calibration, and recording is

necessary.

Design Implications

Valid studies involving psychophysiological measures invariably employ pretest-posttest control group designs. Multiple or continuous measurements are included prior to, during, or following experimental manipulations. The pretest measurements constitute “baseline” or resting levels, allowing a check on desired physiological adaptation to the recording environment itself and a check on physiological comparability of randomly assigned groups. The multiple or continuous within-participant measurement strategy proves advantageous for the indexing strategy already delineated (see propositions presented earlier in the chapter) and allows the investigator to statistically minimize individual differences in basal physiological responses. For more complicated versions of the pretest-posttest designs, counterbalancing strategies minimize adaptation effects for multiple within-subjects manipulations (West, Cham, & Liu, Chapter 4 in this volume).

Useful Physiological Indexes of Psychological Constructs

To this point, we have reviewed information important to psychophysiological indexing, albeit briefly (we encourage interested readers to avail themselves of more of the same on their own). Although this information might prove interesting in its own right, its value would certainly shrink (and this chapter would not be included in this volume) if valid and reliable psychophysiological indexes of critical constructs for testing theory in social psychology did not exist. Fortunately, several important psychophysiological indexes that appear in the literature have been validated and used, and more will surely follow. Although we have chosen here not to exhaust all plausible physiological indexes of psychological constructs of interest to social psychologists here, our illustrations stem not just from the limits of chapter length but also from our judgment that the specific indexes described later assess critical motivational and affective states or processes, are well validated, and familiar to the author.

Psychophysiological Indexes of Motivational States: Challenge and Threat

Do academic performance stereotypes really threaten minority group members as Steele and his colleagues (Steele & Aronson, 1995) have hypothesized? Are low self-esteem individuals challenged by task failure as Swann (1983) has suggested? Do self-protective strategies such as self-handicapping reduce threat as Berglas and Jones (1978) have hypothesized? What coping strategies increase challenge or reduce threat? Does interacting with stigmatized contact promote threat as Goffman (1963) theorized? Confirmation of these (e.g., Blascovich, Mendes, Hunter, Lickel, & Kowai-Bell, 2001) and myriad other long-held hypotheses (see Blascovich, 2008 for a review) has benefited from the utilization of a psychophysiological index of challenge and threat developed more than a decade ago in our laboratory. Here, we review the rationale underlying the index and its validation and provide an example or two of its use.

Rationale

Constructs.

Recall that fuzzy or implicit definitions of many psychological constructs impede the successful development of any indexes of them, including psychophysiological indexes. We can, however, explicitly define the motivational constructs we label as challenge and threat in terms of individuals' relative assessments or evaluations (conscious and nonconscious) of situational demands and available resources to meet those demands (Blascovich, 2008). Accordingly, *challenge* results from the evaluation of resources as meeting or exceeding demands, and *threat* results from the evaluation of demands exceeding resources.

Context.

Recall also that one can develop strong psychophysiological indexes of a construct by narrowing or limiting the psychosocial context. For example, one cannot necessarily interpret increases in cardiac performance such as ventricular contractility (i.e., the strength of the pumping action of the left ventricle) and cardiac output (i.e., blood flow) that occur in metabolically demanding situations such as jogging in the same way that one interprets increases that occur in situations without such metabolic demands. Likewise, even within nonmetabolically demanding situations, one must evaluate increases in cardiac performance in situations requiring active cognitive responses differently than in

those situations requiring passive endurance (Blascovich, 2008; Obrist, 1981).

Of course, limiting the psychosocial context of psychophysiological indexes precludes the possibility of invariant indexes. However, this downside is not particularly problematic, as invariance can be claimed for relatively few indexes in psychology and because limiting the generality does not mean rendering the index meaningless. Provided the context represents a large or meaningful cross-section of the behavioral domain, one can aim for a marker, thereby providing the basis for strong inference within the specified type of context.

In the development of the cardiovascular indexes of challenge and threat, we limited the context to what we label *motivated performance situations*. Motivated performance situations are goal-relevant to the individual, thereby engendering some degree of task engagement. Furthermore, motivated performance situations require instrumental cognitive responses and often overt actions on the part of the performer. We further limited motivated performance situations to nonmetabolically demanding ones – that is, ones excluding gross, repetitive large-muscle movements. Although these limits rule out many social psychologically relevant situations, such as athletic competitions, they rule in a very important and large behavioral domain. Taking academic and other examinations, preparing and giving speeches, conducting interpersonal negotiations, making decisions and judgments, initiating close relationships, and interviewing for a job all qualify as nonmetabolically demanding motivated performance situations. Individuals encounter these situations every day, and many are quite goal relevant for them. Furthermore, social psychologists have traditionally utilized motivated performance tasks in a wide variety of experimental contexts, testing an even wider variety of theories.

One-To-One Relationships.

Even though challenge and threat clearly represent different motivational states, each of which likely involves different physiological responses, in the last two decades social psychologists have focused on differences between physiological markers of these motivational states in what we have come to label as motivated performance situations as Seery (2011) has reviewed. The failure to distinguish between different patterns of cardiovascular responses associated with positive and negative motivational states in the past propelled both a large literature connecting cardiovascular performance increases, as indexed conveniently but perhaps naïvely by unitary measures such as heart rate changes and blood pressure increases, to the negative consequences of stress or threat on the

cardiovascular system (e.g., Blascovich & Katkin, 1993), and a separate literature using virtually the same indexes connecting cardiovascular increases to the positive consequences of motivation and positive performance (e.g., Blascovich, Seery, Mugridge, Weisbuch, & Norris, 2004; Seery, Weisbuch, Hentenyi, & Norris, 2010).

Thus, cardiovascular changes such as increases in heart rate and blood pressure were used by different researchers to index oppositional motivational states, creating a dilemma for those interested in using cardiovascular indexes to identify distinctively one or the other superordinate motivational states (e.g., challenge or threat). This dilemma stems, of course, from the many-to-one relationships that hold for measures such as heart rate. Fortunately, however, the dilemma has been recognized and dealt with at theoretical and methodological levels.

Working within the tradition of cardiovascular reactivity and psychological stress, Dienstbier (1989) challenged the view that increased cardiovascular performance during potentially stressful situations is necessarily associated with malignant psychological states and, on the basis of both human and animal research, theorized that increases in cardiovascular performance could be and are often associated with positive or nonmalignant states. Dienstbier posited that increased sympathetic-adrenomedullary (SAM) activity is associated with benign states and improved performance and that increased pituitary-adrenocortical (PAC) activity is associated with malignant states when such activation occurs alone or concomitant with SAM activation.

Interestingly, about the same time as Dienstbier's article appeared, Kasprovicz and his colleagues (Kasprovicz, Manuck, Malkoff, & Krantz, 1990) published a piece categorizing individuals, based on the preponderance of specific types of cardiovascular responses, as either “cardiac” or “vascular” responders. The former respond primarily with changes in activity in the heart, and the latter with changes in activity in the arteries. Later, Manuck, Kamarck, Kasprovicz, and Waldstein (1993) suggested that so-called vascular reactivity appeared to be the more pernicious of the two, heart-health-wise.

Encouraged by the theoretical work of Dienstbier (1989) and the empirical work of Kasprovicz *et al.* (1990), and keeping in mind that we wished to ensure the viability of one-to-one relationships between selected physiological measures and challenge and threat motivational states, we applied Dienstbier's theoretical rationale to the selection and collection of a set of cardiovascular responses of human participants in motivated-performance situations. Based on the work of

Dienstbier (1989) as well as on the work of Gray (1982) and McNaughton (1993), we posited that the benign pattern of physiological activation marked by SAM activation caused (a) sympathetic neural stimulation of the myocardium increasing cardiac performance and (b) adrenal medullary release of epinephrine causing vasodilation in the large muscle beds and lungs and an overall decrease in systemic vascular resistance, as well as some additional enhancement of cardiac performance. We posited that the malignant pattern marked by dual activation of the PAC and SAM axes caused (a) elevations of cardiac performance over resting levels (SAM activity), and (b) decreased release of epinephrine and norepinephrine from the adrenal medulla (PAC activity), causing moderate increases in cardiac output without accompanying decreases in systemic vascular resistance.

Logically, then, we needed a set of measures including separate unambiguous measures of both cardiac and vascular performance. And, of course, we needed to be able to assess these measures practically in a technological sense. Because of their ambiguity in terms of cardiac and vascular underpinnings, simple heart rate and blood pressure measures could not be used as unambiguous measures. Indeed, their ambiguity led to the indexing dilemma described earlier in the first place. Fortunately, the use of impedance cardiography had emerged in psychophysiology (Sherwood et al., 1990), enabling researchers to assess less ambiguous measures of cardiac performance, such as pre-ejection period (PEP), a measure of ventricular contractility, and cardiac output (CO), a measure of blood flow. In addition, continuous blood pressure monitoring became available, which together with the impedance-derived measures allowed the noninvasive assessment and calculation of total peripheral resistance (TPR), an unambiguous measure of vascular performance.

According to Dienstbier's (1989) physiological toughness theory, challenge and threat could be indexed by different patterns of cardiac and vascular responses over time. Specifically, we expected that challenged humans, those for whom resources (i.e., skills, abilities, etc.) outweighed the demand, (i.e., danger, uncertainty, and required effort) in a motivated-performance situation, should show relatively large increases in cardiac performance (as indexed by PEP and heart rate; HR), increases in blood flow (as indexed by CO), and relatively large decreases in vascular resistance (as indexed by TPR). We expected that threatened individuals, those for whom demands in a motivated performance situation outweighed resources, should also show increases in cardiac performance (as indexed by PEP, HR), little change or decreases in blood flow (as indexed by CO), and no change or increases in vascular resistance (as

indexed by TPR). Figure 6.4 depicts the predicted changes as change or difference scores from rest to task performance.

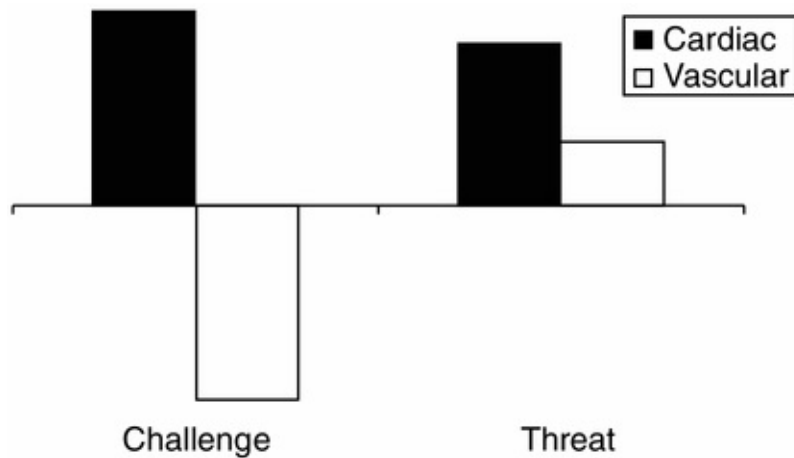


Figure 6.4. Predicted cardiovascular markers of challenge and threat (change scores).

Validational Research

To validate these indexes, we conducted three types of studies: correlational, experimental, and manipulated physiology. We wanted to know if our predicted patterns of cardiovascular responses (see Figure 6.4) were associated with free evaluations of challenge and threat, if we could evoke the patterns by manipulating the motivated performance situation in ways likely to cause challenge and threat motivational states, and whether the psychological states drive the cardiovascular responses or vice versa.

Correlational Studies.

These studies involved an initial one as well as a cross-sectional replication (Experiments 2 and 3 in Tomaka, Blascovich, Kelsey, & Leitten, 1993). In these studies, participants in an experimentally created motivated-performance situation received instructions regarding an upcoming mental arithmetic task requiring quick and accurate vocal serial subtractions (e.g. “7s,” “13s,” etc. from a four-digit number). After receiving instructions, but prior to actual task performance, self-reported demand and ability evaluations were solicited from participants, allowing us to assess evaluations overall by calculating demand-resource ratios. On the basis of these ratios, we were able to divide participants into challenge and threat groups. Subsequent analyses in both studies revealed the same patterns of cardiovascular responses associated with free challenge and

threat evaluations as those predicted earlier. Specifically, as illustrated in [Figure 6.5](#), in both studies challenged participants exhibited the benign pattern of cardiovascular responses described previously, including relatively large increases in cardiac activity accompanied by decreases in total systemic vascular resistance. Threatened participants exhibited the less benign pattern, including increases in cardiac activity and increases in total systemic vascular resistance.

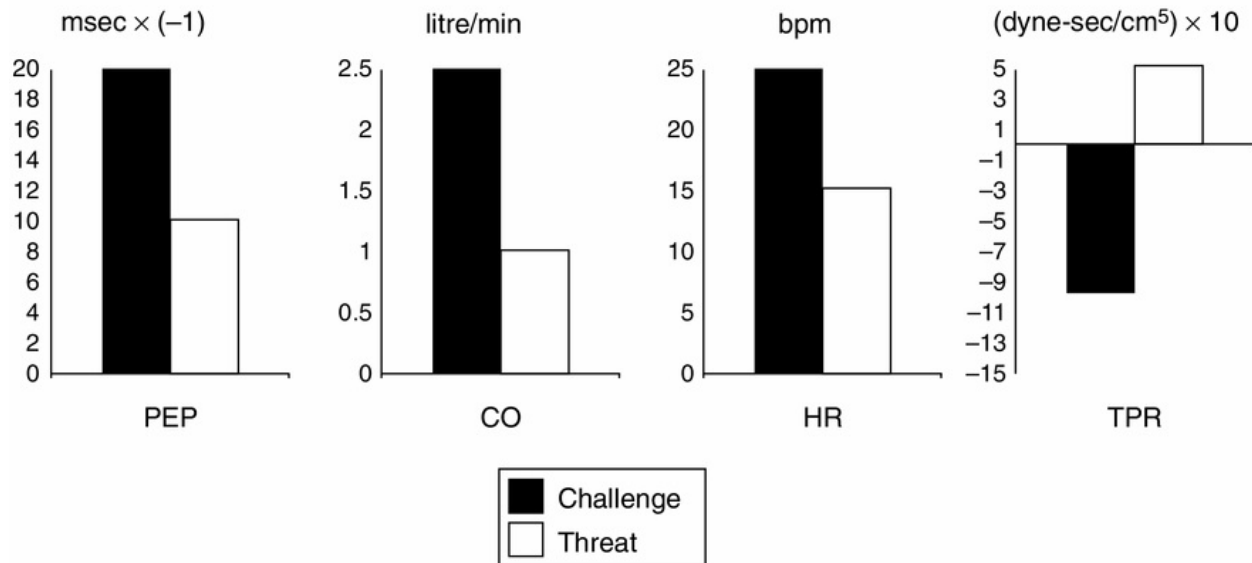


Figure 6.5. Free-appraisal study 1 results.

In addition, challenged participants in both studies recalled less stress during their task performance, perceived greater effort, and perceived better performance than threatened participants.

Experimental Studies.

Because these studies were essentially correlational in nature, we needed to determine if the challenge and threat patterns of cardiovascular response could be evoked by experimental manipulation (Experiment 1 in Tomaka, Blascovich, Kibler, & Ernst, 1997). Again, participants performed a mental arithmetic task, in this experiment after receiving an instructional set emphasizing either threat (highlighting accuracy of task performance and potential evaluation) or challenge (highlighting effort and doing one's best). Cognitive evaluations were assessed after instructions but prior to performing the task. Physiological responses were recorded continuously during the task and during a quiet rest period preceding the task.

Analyses indicated that the instructional set had the expected effects on

demand/resource ratios such that participants receiving instructions emphasizing threat showed greater demand-resource ratios than participants receiving instructions emphasizing challenge. Furthermore, those receiving challenge instructions exhibited the benign pattern of cardiovascular responses and those receiving threat instructions exhibited the less benign and predicted pattern.

Manipulated Physiology Studies.

Although clearly supporting the logic of our cognitive appraisal approach to the validation of cardiovascular indexes of challenge and threat motivation in motivated performance situations, the results of the studies described in the preceding section do not exclude the possibility that physiological responses drive the psychological responses. To test the latter notion, distinct patterns of autonomic physiological activity consistent with threat and challenge were evoked nonpsychologically, and the resulting effects of such manipulations on evaluations were examined (Experiments 2 and 3 in Tomaka et al., [1997](#)).

Two different patterns of physiological activation, each having a distinct physical mode of elicitation, were employed. The first was a manipulation of cardiovascular reactivity consistent with challenge. For this manipulation, participants pedaled a stationary bike for a relatively short period of time, but long enough to achieve relatively high cardiac reactivity coupled with a decline in systemic vascular resistance. The second was a manipulation of vascular reactivity consistent with threat accomplished by exposing participants to the cold pressor task. Cold pressor tasks have been shown to produce vasoconstrictive responses and increases in systemic vascular resistance (Allen, Shelley, & Boquet, [1992](#)). Checks revealed that the nonpsychological manipulations produced appropriate challenge and threat-like patterns of cardiovascular responses. However, no effects were found on overall challenge and threat evaluations.

Predictive Validation Studies.

In nearly every challenge-threat study conducted by ourselves and others, participants' task performance (serial subtraction, word finding, and speeches) during experiments was better in a challenge compared to threat state (as indexed by the pattern of cardiovascular measures they exhibited). Consequently, we conducted predictive validation studies to determine if the physiological indexes were predictive of future performance.

In the first (Blascovich, Seery, Mugridge, Weisbuch & Norris, 2004), college varsity baseball and softball players gave two short speeches: a baseball-relevant speech (“How I would approach a critical hitting situation”) and a baseball-irrelevant speech (“Why I am a good friend”). We derived a unitary cardiovascular index from the challenge-threat-relevant cardiovascular measures for each participant. Controlling for the cardiovascular index during the baseball-irrelevant speech, the cardiovascular index during the baseball-relevant speech reliably predicted players’ offensive baseball performance six months later during the varsity baseball and softball seasons. In a second study (Seery, Weisbuch, Hetenyi, & Blascovich, 2010), participants gave a specific undergraduate course–relevant speech (i.e., a course they had just started) or the friend speech. Again, the derived index controlling for the course-irrelevant speech reliably predicted students’ performance in the course.

Summary of Validation Research.

These studies affirmed our notion that theoretically (i.e., from Dienstbier, 1989) derived and distinctively different patterns of cardiovascular responses accompany challenge and threat motivation in nonmetabolically demanding motivated-performance situations. The distinctive challenge (i.e., increases in cardiac performance, increases in blood flow, and decreases in vascular resistance) and threat (i.e., increases in cardiac performance, little or no increase in blood flow, and increases in vascular resistance) held in correlational, experimental, and predictive validation studies involving both freely generated and manipulated evaluations related to challenge and threat motivation. Nonpsychological manipulations of the different patterns of cardiovascular responses themselves did not produce differences in psychological motivational states, affirming the theoretical rationale underlying the indexes.

Research Examples

Belief in a Just World.

The availability of the cardiovascular indexes of threat and challenge enabled us to examine dispositional influences on evaluation processes in motivated performance situations. One disposition we (Tomaka & Blascovich, 1994) explored is “belief in a just world” (BJW), which is the extent to which individuals believe that people generally “get what they deserve” from life or

conversely the extent to which individuals believe that “life is inherently unfair” (Lerner, 1980). According to several theorists (e.g., Lazarus & Folkman, 1984; Lerner, 1980; Lerner & Miller, 1978), dispositional belief in a just world protects individuals, allowing them to adapt better to the demands of everyday life. In a motivated-performance situation, high-BJW individuals should exhibit challenge motivation in contrast to low-BJW individuals.

In our study, which involved a motivated-performance situation incorporating mental arithmetic, we blocked participants on dispositional BJW. As expected, task evaluations, cardiovascular response patterns, and performance differed as a function of BJW. High-BJW participants made more challenging overall evaluations, exhibited the challenge pattern (i.e., strong increases in cardiac performance coupled with decreased peripheral resistance), and performed better than low-BJW participants, who exhibited the threat pattern (i.e., increases in cardiac performance coupled with slight vasoconstriction).

Attitude Functionality.

Functionally, attitudes should facilitate decision making (Allport, 1935; Fazio, 1989; Katz, 1960) by providing individuals with relatively accessible knowledge enabling them to make decisions in demanding situations more easily. Task object-relevant attitudes should increase the probability of a challenge rather than threat in a motivated-performance situation. We (Blascovich et al., 1993; Experiment 2) explored these issues in a two-phase experiment. In the first phase, participants developed attitudes toward sets of novel objects (abstract paintings) using Fazio's procedure (e.g., Fazio, Chen, McDonel, & Sherman, 1982). Half of the participants (15) rehearsed attitudes toward one subset of the abstract paintings, and the other half (15) toward a second mutually exclusive subset. In the second phase, a motivated-performance situation, participants expressed rapid pairwise preferences for 34 slides of randomly paired abstract paintings (i.e., attitude objects). Half of each participant group vocalized preferences within paired abstract paintings selected from the subset toward which they had rehearsed attitudes, whereas the other half vocalized preferences within pairs selected from the unfamiliar subset. Participants in the rehearsed-painting condition exhibited increased cardiac response and vasodilation, the challenge pattern, whereas those in the novel-painting condition exhibited increased cardiac response and vasoconstriction, the threat pattern.

Stigma.

Much theoretical work starting with Goffman (1963) hypothesized that non-stigmatized individuals are threatened during social interactions with stigmatized (i.e., by race, socioeconomic status, physical deformities, etc.) individuals. However, little if any research had confirmed this hypothesis. In a series of studies (e.g., Blascovich, Mendes, Hunter, Lickel, & Kowai-Bell, 2001; Mendes, Blascovich, Lickel, & Hunter, 2002), utilizing the challenge-threat indexes, we demonstrated that non-stigmatized individuals were indeed threatened during interactions with stigmatized individuals such as individuals bearing facial birthmarks, African Americans, or individuals with low socioeconomic status.

A Word on Technology.

The assessment of the cardiovascular indexes of challenge and threat discussed in the preceding sections requires incorporating impedance cardiographic and blood pressure recording devices into social psychology research, which, historically, has been limited to use in an actual physical laboratory. Fortunately, these devices are not only reliable and relatively self-contained but today are also physically small, about the size of a trade-sized paperback book and a DVD player, respectively, and even smaller wireless devices are currently available for ambulatory use outside of the laboratory. These devices perform the transducing and conditioning processes in the signal path described earlier. Turnkey software is also available, allowing direct recording of continuous cardiac (e.g., PEP, SV, CO) and blood pressure responses and the calculation of vascular response (i.e., total peripheral resistance). Technical details regarding the placement of sensors, operation of the equipment, and software are beyond the scope of this chapter and are available in several sources, including Kelsey and Guethlein (1990), Sherwood (1993), and Sherwood *et al.* (1990).

Psychophysiological Indexes of Affective States: Positive and Negative Affect

Testing of theoretical hypotheses related to affect can also benefit from the use of psychophysiological indexes. Here, we review two such indexes: facial electromyography and startle eyeblink reflex responses. First, however, we attend to some general issues regarding the psychophysiological indexing of affect.

Constructs.

In the past, explicit definitions of positive and negative affect proved no less daunting than definitions of motivational constructs such as challenge and threat, although arguably more theoreticians agree on definitions of positive and negative affect than on the definitions of moods or specific emotions. Basically, positive and negative affects represent superordinate emotion or feeling categories (cf. Feldman-Barrett, 2006; Russell, 2003 ; Shaver, Schwartz, Kirson, & O'Connor, 1987) encompassing several specific basic and many subordinate emotions. Phenomenologically, positive affect occurs during the experience of any specific positive mood or emotion, and negative affect occurs during the experience of any specific negative mood or emotion. Most indexes, nonphysiological as well as physiological, of affect rest on the idea of commonalities among the specific positive emotions as a group and something else among the specific negative emotions as a group.

Context.

In the development of physiological indexes of positive and negative affect, researchers have generally limited the empirical context to situations requiring relatively little movement and involving passive receipt of stimuli or information by individuals. These empirical limits, however, do not preclude the extension of physiological indexes of affect to more psychologically and physically active situations, such as motivated performance situations, at least theoretically. However, for physiological indexes of affect, which involve small muscle assessments (see discussion later in the chapter), practical aspects of physiological recording (i.e., secure attachment of sensor and high signal/noise ratios) generally limit research to contexts involving relatively little physical movement.

One-To-One Relationships.

Positive and negative affects represent different feeling states, each of which according to the identity thesis (see earlier discussion) presumably involves some physiological response. Cacioppo, Petty, Losch, and Kim (1986) and Lang, Bradley, and Cuthbert (1990) challenged the validity of physiological indexes of affect valence such as skin conductance or heart rate changes because of a lack of one-to-one relationship with affect constructs (albeit leaving open the possibility that these unitary physiological indexes might relate to the intensity but not the valence of affect). As researchers have sought more sophisticated psychophysiological indexes, those with one-to-one relationships to affective

states, they have focused on patterns of facial electromyographic (EMG) recordings specific to the facial muscles associated with the expression of affect, and on the inhibition-facilitation of reflexes (i.e., startle eyeblink responses) during the experience of positive and negative affective states.

Facial EMG Indexes of Positive and Negative Affect

Rationale.

Cacioppo and his colleagues (e.g., Cacioppo & Petty, 1981; Cacioppo, Petty, & Marshall-Goodell, 1984; Cacioppo et al., 1986) proposed the value of using electromyograms specific to targeted facial muscles as physiological indexes of positive and negative affect. They based their rationale on long-lived scholarly interest in the muscles of facial expression beginning in the modern era with the work of Darwin (1965 [1872]) on the evolutionary significance of facial expressions to the more recent facial expression work of Ekman and his colleagues (e.g., Ekman, 1993). Cacioppo's group reasoned that because the “somatic nervous system is the final pathway via which people interact with and modify their physical and social environments” (Cacioppo et al., 1986, p. 261) and particularly because the face is the locus of most emotional expression, targeted facial muscle responses represented fertile ground for psychophysiological indexing of affect.

Cacioppo's group focused in particular on the corrugator supercilii (i.e., “frown muscles”) and the zygomaticus majori (i.e., “smile muscles”). Figure 6.6 depicts these and other facial muscle sites. This focus derived from earlier work (e.g., Fridlund, Schwartz, & Fowler, 1984) demonstrating that zygomaticus and corrugator EMG activity varied with participants' emotional reactions to videotapes and during emotional imagery states. Cacioppo and his colleagues were also convinced that the highly sensitive EMG recordings could detect covert movement in the targeted facial muscle groups and reveal affectively meaningful psychological states not detectable by visual observation. This latter point becomes important as individuals can exert control over overt facial expressions, leading to a disparity between overt expressions and covert affective states.

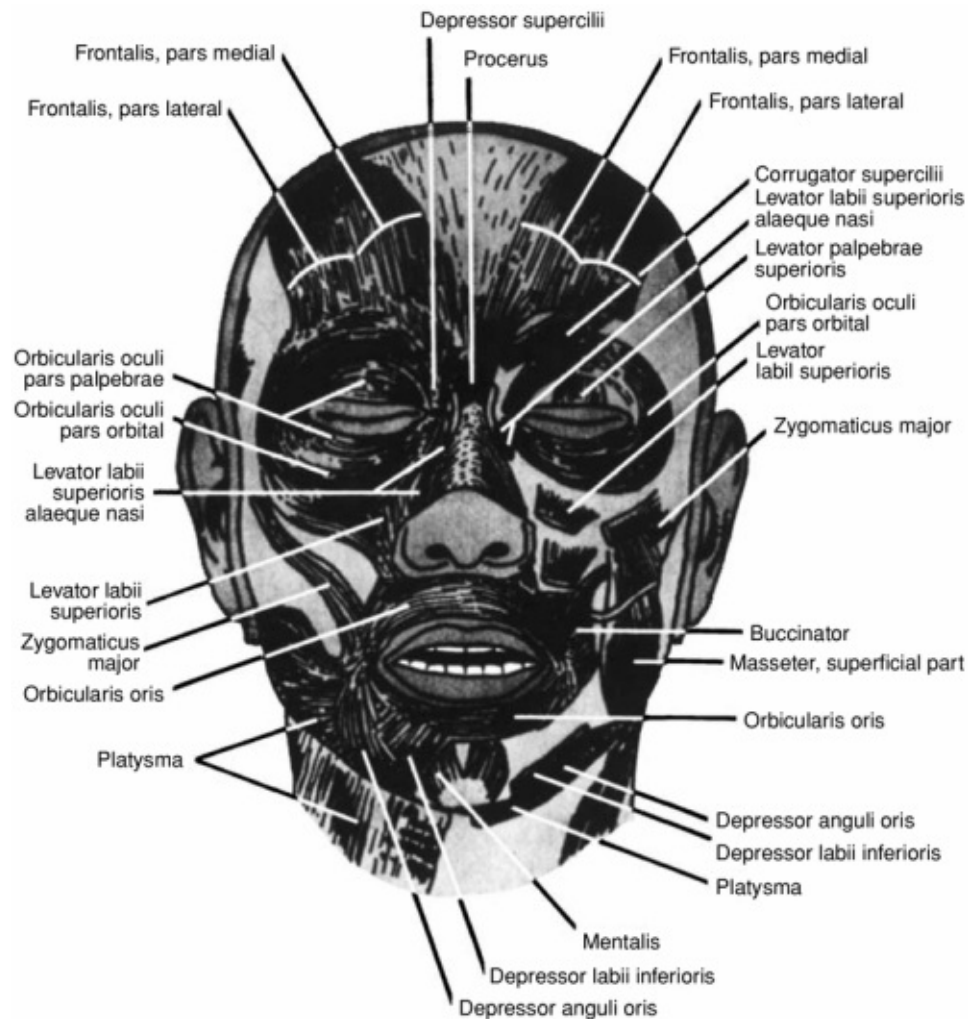


Figure 6.6. Facial muscles.

Figure 6.7 depicts Cacioppo et al.'s (1986) description of the basic patterns of facial EMG responses marking positive and negative affect. Accordingly corrugator supercilii EMG increases and zygomaticus majori EMG decreases during negative affect, whereas corrugator supercilii EMG decreases and zygomaticus majori EMG increases during positive affect.

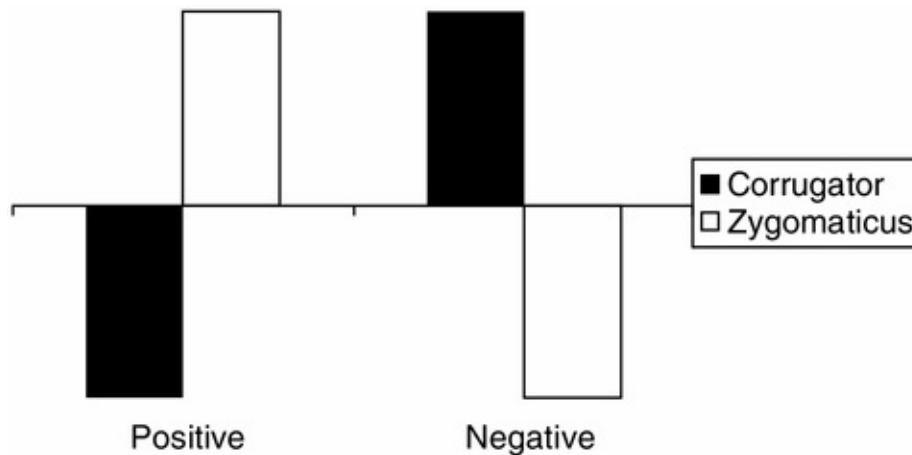


Figure 6.7. Predicted facial electromyographic markers of positive and negative affect.

Validational Research.

Cacioppo *et al.* (1986) reported two very similar studies validating their proposed EMG indexes of affect. In the major study reported, the investigators presented 28 participants with affectively valenced (positive and negative) visual stimuli, both mild and moderate in intensity, as well as neutral and buffer stimuli. The investigators recorded facial and nonfacial EMG during each stimulus presentation (5 sec), including muscles hypothetically related (e.g., corrugator supercilii and zygomaticus majori) and unrelated (e.g., superficial forearm flexor, medial frontalis) to affect. The researchers gathered self-reports of liking, arousal, and familiarity following each stimulus presentation. Basically, Cacioppo *et al.* (1986) confirmed the utility of the proposed EMG indexes of affect. Zygomaticus majori and corrugator supercilii EMG varied as a function of valence, with the former increasing during positively valenced stimuli and the latter increasing during negatively valenced stimuli. The investigators found no differences among a set of hypothetically nonrelated muscles (i.e., orbicularis oris, medial frontalis, and superficial forearm flexor), thus ruling out any contribution of general muscle tension to the indexes.

Cacioppo *et al.* (1986) not only tested for predictable relationships between facial EMG markers and affective states but also between potentially human-detectable visual markers (i.e., facial expressions) and affective states. Human judges simply were unable to distinguish between positive and negative affective states based on videotapes made during physiological recording, most likely because the carefully chosen stimuli elicited minimal levels of affect. Both the validation of the EMG markers and the lack of same for the visual markers

caused Cacioppo *et al.* (1986) to conclude that “facial EMG can mark the valence and intensity of transient and specific affective reactions even in the absence of emotional expressions that are noticeable, at least under normal viewing conditions” (p. 267).

Research Examples: Prejudice and Discrimination

The use of self-report measures in the domain of prejudice and discrimination has long proven problematic for obvious reasons of self-presentation. However, the use of electromyography in this domain has proven more effective. Vanman and colleagues, for example, applied the use of facial electromyography in this domain.

For example, Vanman, Paul, Ito, and Miller (1997) found and replicated effects such that while participants self-reported more positive affect for black than white partners, they exhibited EMG activity (i.e., increased corrugator supercilii activity and lower zygomaticus major activity) indicative of more negative affect for black than white partners. Vanman and colleagues also demonstrated that modern racism scores were related in expected ways to facial corrugator and zygomaticus EMG activity during observation of black and white target stimulus photos, but not their overt friendliness ratings of the target photos. In terms of predictive validity, Vanman, Saltz, Nathan and Warren (2004) linked negative affect revealed via facial EMG to racial bias in a personnel-hiring scenario.

Technology.

The assessment of the facial somatic indexes of positive and negative affect requires EMG technology. Such technology covers the entire signal path described earlier. Technical details regarding the placement of sensors, operation of the equipment, and software are beyond the scope of this chapter but are available in several sources, including Blascovich *et al.* (2011).

Startle Eyeblink Reflex Indexing of Positive and Negative Affect

Rationale.

Lang and his colleagues (Lang, Bradley, & Cuthbert, 1990, 1992) proposed the

value of using electromyograms specific to reflexive eye blinks to index affective valence. Guided by the work of Schneirla (1959), Konorski (1967), Dickinson and Dearing (1979), and Masterson and Crawford (1982), Lang *et al.* (1990, 1992) based their rationale on the assumption that brain states organize behavior along an appetitive-aversive dimension. They postulated that positive affect is associated with a brain state favoring approach, attachment, and consummatory behavior, and that negative affect is associated with a brain state favoring avoidance, escape, and defense. They argued further that “[t]he efferent system as a whole (including exteroceptive reflexes) is presumably tuned according to the current states of this central affect-motivational organization” (p. 377).

Accordingly, Lang *et al.* (1990) hypothesized that reflexes associated with positive affect would be enhanced during the ongoing experience of a positive emotional state and that reflexes associated with negative affect would be enhanced during the ongoing experience of a negative emotional state. Furthermore, they hypothesized that affectively valenced reflexes would be inhibited during the ongoing experience of the affectively opposite emotional state. Thus, changes (i.e., facilitation or inhibition) in affectively toned reflexes could be used to index the basic ongoing affective state of an individual.

Lang and his colleagues focused on the startle eyeblink reflex, the reflexive blinks that occur when individuals perceive an unexpected and relatively intense stimulus, particularly –, but not limited to – an acoustic stimulus. The startle eyeblink reflex is negatively toned and, hence, should be enhanced during ongoing negative affect and inhibited during ongoing positive affect. The eyeblink is also relatively easy to measure physiologically using orbicularis oculi EMG.

Validational Research.

In an elegant series of studies (see Lang *et al.*, 1990 for a review), Lang and his colleagues provided convincing evidence for the startle eyeblink index of ongoing affect. Vrana, Spence, and Lang (1988) tested the hypothesis that the valence of ongoing affect enhances or inhibits acoustically driven startle eyeblink reflexes according to the match or mismatch between the valence of the underlying affective state and this negatively toned reflex. These investigators presented 36 negative, neutral, and positive photographs via slides to participants for 6-second periods during which they presented unpredictable, loud white-noise bursts binaurally to participants while using orbicularis oculi EMG to

measure the strength of reflexive eyeblink responses. As predicted, the data supported their hypothesis. A conceptual replication and extension by Bradley, Cuthbert, and Lang (1988) produced the same pattern of data, this time allowing participants to control the length and duration of stimulus slide presentation. Bradley, Cuthbert, and Lang (1990) produced the same effects substituting visually evoked startle eyeblink reflexes proving the eyeblink reflex effects were independent of startle stimulus modality.

Vrana and Lang (1990) demonstrated that the startle reflex methodology indexed affective state during affectively relevant imagery and memories. This study provided the crucial evidence that the startle reflex methodology is sensitive to internally generated affective states – a necessary assumption if social psychological investigators are to benefit from the techniques of physiological indexing of covert affect suggested by Lang and his colleagues.

Startle Reflex Responses and Social Psychological Research

Although the startle eyeblink reflex measure of Lang and his colleagues has enjoyed increasing popularity among psychophysicists, this index has attracted the attention of social psychologists more slowly. Nevertheless, we urge researchers to consider it, as the startle eyeblink reflex index is relatively simple to employ and its validity is based on a wealth of research.

Technology.

The assessment of the startle eyeblink reflex index of positive and negative affect generally requires EMG technology. As with facial EMG measures of affect described earlier, such technology covers the entire signal path previously discussed. Technical details regarding the placement of sensors, operation of the equipment, and software are beyond the scope of this chapter but are available in several sources, including Cacioppo, Tassinary, and Fridlund (1990) and Lang *et al.* (1990).

Other Physiological Indexes of Psychological Constructs Important to Social Psychologists

Although the limits of this chapter preclude detailed illustrations of additional physiological indexes of psychological constructs, the reader should feel no such

constraints. Indeed, other valid psychophysiological indexes exist and more continue to appear as psychophysiological theory and technology advances. Notably, promising physiological indexes of cognitive processes exist (see Heinze, Munte, & Mangun, 1994 for a review). For example, orbicularis oris EMG activity has been proposed to index verbal processing (Cacioppo & Petty, 1981). Startle reflex techniques have been used to index attention (Anthony & Graham, 1985). Event-related potentials have been investigated as indexes of memory processes (Nielsen-Bohlman & Knight, 1994).

Assessment and Summary

As we predicted more than a decade ago (Blascovich, 2000), steady growth has occurred in the use of physiological indexes in social psychology as increasingly more researchers recognize their unique benefits and avail themselves of necessary biological and methodological background opportunities. Physiological indexes will not completely replace more traditional self-report and behavioral indexes in social psychology. However, we dare say that as physiological indexes have been more properly employed and become more widespread, they are impacting social psychological research more substantially. Here we summarize the general points we have made in this chapter in terms of brief answers to a number of general questions.

How Should Interested Social Psychologists Begin the Process of Implementing Psychophysiological Indexes?

Researchers must think about using physiological indexes in the same ways they think of using subjective and behavioral indexes, convincing themselves and others of the validity and reliability of such indexes as markers and invariants (i.e., bearing one-to-one relationships) of the psychological constructs under scrutiny. Specific psychophysiological indexes derive their validity from psychophysiological theory confirmed via systematic empirical work. Furthermore, we have noted that derivation of successful physiological indexes, like nonphysiological indexes, largely depends on conceptual explication of target psychological constructs, on identification of key physiological response patterns across multiple physiological measures, and on specification of appropriate situational contexts.

What Sorts of Ideas Are Better or Worse Served by Psychophysiological Indexes?

Physiological indexes lend themselves to assessment of continuously fluctuating psychological states and processes, especially affective, motivational, and cognitive ones. Thus, physiological responses are unlikely to usefully index specific attitudes, beliefs, and dispositions. However, to the extent that specific attitudes, beliefs, and dispositions influence affective, motivational, and cognitive processes, physiological indexes of the latter processes can be used effectively to test theoretical arguments and hypotheses regarding specific constructs.

What Advantages Accrue to Physiological Indexes?

Because physiological indexes are objective and covert, avoidance of self-report bias presents one of the major advantages of physiological indexes. However, because physiological indexes are also continuous, other advantages accrue. Thus, one can assess psychological states and processes continuously over time providing data for powerful time-series analytic statistical techniques. More importantly, one can assess psychological states and processes in the background concurrently with other types of measures including self-report and behavioral ones or while research participants engage in other activities. Thus, physiological measures do not interfere with experimental treatments, and such indexes can eliminate the need for post hoc or retrospective accounts of target states and processes.

Do Physiological Indexes Provide the “Gold Standard” for Psychological Measurement?

That physiological measures enjoy intrinsic superiority over other types is a naïve notion that perhaps stems from the mystique of physiological measurement discussed at the outset of this chapter. Rather, physiological indexes provide a third set of measurement methods in addition to subjective and behavioral ones. Together with the other types of measures, physiological ones add to the power of multimethod triangulation as indicated earlier in this chapter and as Brewer and Crano (Chapter 2) describe in this volume.

Where Can One Find a Catalog of Valid Physiological

Indexes of Psychological Constructs?

In theory, a catalog of psychophysiological indexes would be pragmatic. We have made an attempt at such elsewhere (Blascovich et al., 2011). Here we did not attempt to provide exhaustive coverage of physiological indexes. Rather, we tried to impart principles and provide illustrations of what we believe are some valid indexes of superordinate psychological constructs. We hope that readers understand the epistemological principles we have presented regarding the development of psychophysiological indexes, and that they are able to judge the validity of proposed physiological indexes in the literature and their application to the assessment of social psychological constructs. Used validly (e.g., incorporating the principles suggested previously), physiological measures provide convincing and fairly unobtrusive evidence to support testable hypotheses and in turn the theories from which such hypotheses are drawn.

References

- Allen, M. T., Shelley, K. S., Boquet, A. J (1992). A comparison of cardiovascular and autonomic adjustments to three types of cold stimulation tasks. *International Journal of Psychophysiology*, 13, 59–69.
- Allport, G. W. (1935). *Attitudes*. In C. Murchison (Ed.), *Handbook of social psychology* (pp. 798–884). Worcester, MA: Clark University Press.
- Anthony, B. J., & Graham, F. K. (1985). Blink reflex modification by selective attention: Evidence for the modulation of ‘autonomic’ processing. *Biological Psychology*, 21, 43–59.
- Berglas, S., & Jones, E. E. (1978). Drug choice as a self-handicapping strategy in response to noncontingent success. *Journal of Personality and Social Psychology*, 36, 405–417.
- Blascovich, J. (2000). Psychophysiological methods. In H.R. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 117–137). Cambridge: Cambridge University Press.
- Blascovich, J. (2008). Challenge and threat. In A. J. Elliot (Ed.), *Handbook of approach and avoidance motivation* (pp. 431–446). New York: Erlbaum.
- Blascovich, J., Ernst, J. M., Tomaka, J., Kelsey, R. M., Salomon, K. A., & Fazio, R. H. (1993). Attitude as a moderator of autonomic reactivity. *Journal of Personality and Social Psychology*, 64, 165–176.

- Blascovich, J., & Ginsburg, G. P. (1978). Conceptual analysis of risk taking in “risky shift” research. *Journal for the Theory of Social Behavior*, 8, 217–230.
- Blascovich J., & Katkin, E. S. (Eds.). (1993). *Cardiovascular reactivity to psychological stress and disease*. Washington, DC: American Psychological Association.
- Blascovich, J., & Kelsey, R. M. (1990). Using cardiovascular and electrodermal measures of arousal in social psychological research. *Review of Personality and Social Psychology*, 11, 45–73.
- Blascovich, J., & Mendes, W. B. (2010). Social psychophysiology and embodiment. In Gilbert, D., Fiske, S., & Lindzey, G. (Eds.), *Handbook of Social Psychology* (5th ed., pp. 194–227): New York: Wiley.
- Blascovich, J., Mendes, W. B., Hunter, S.B., & Lickel, B., & Kowai-Bell, N. (2001). Perceiver threat in social interactions with stigmatized others. *Journal of Personality and Social Psychology*, 80, 253–267.
- Blascovich, J., Seery, M., Mugridge, C., Weisbuch, M., & Norris, K. (2004). Predicting athletic performance from cardiovascular indicators of challenge and threat. *Journal of Experimental Social Psychology*, 40, 683–688.
- Blascovich, J., & Tomaka, J. (1996). The biopsychosocial model of arousal regulation. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 28, pp. 1–51). New York: Academic Press.
- Blascovich, J., Vanman, E. J., Mendes, W. B., & Dickerson, S. (2011). *Social psychophysiology for social and personality psychology*. London: Sage.
- Boyd, R.W., & DeMascio, A. (1954). Social behavior and autonomic physiology: A sociophysiological study. *Journal of Nervous and Mental Disease*, 120, 207–212.
- Bradley, M. M., Cuthbert, B. N., & Lang, P. J. (1988). Lateral presentation of acoustic startle stimuli in a varying affective foreground [abstract]. *Psychophysiology*, 25, 436.
- Bradley, M. M., Cuthbert, B. N., & Lang, P. J. (1990). Startle reflex modification: Attention or emotion? *Psychophysiology*, 27, 513–522.
- Cacioppo, J. T., & Petty, R. E. (1981). Electromyograms as measures of extent and affectivity of information processing. *American Psychologist*, 36, 441–

456.

- Cacioppo, J. T., & Petty, R. E. (Eds.). (1983). *Social psychophysiology: A sourcebook*. New York: Guilford Press.
- Cacioppo, J. T., Petty, R. E., Losch, M. E., & Kim, H. S. (1986). Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of Personality and Social Psychology*, 50, 260–268.
- Cacioppo, J. T., Petty, R. E., & Marshall-Goodell, B. (1984). Electromyographic specificity during simple physical and attitudinal tasks: Location and topographical features of integrated EMG responses. *Biological Psychology*, 18, 85–121.
- Cacioppo, J. T., & Tassinary, L. G. (1990a). Inferring psychological significance from physiological signals. *American Psychologist*, 45, 16–28.
- Cacioppo, J. T., & Tassinary, L. G. (1990b). Psychophysiology and psychophysiological inference. In J. T. Cacioppo & L. G. Tassinary (Eds.), *Principles of psychophysiology: Physical, social, and inferential elements* (pp. 3–33). New York: Cambridge University Press.
- Cacioppo, J. T., Tassinary, L. G., & Berntson, G. (Eds.). (2007). *Handbook of psychophysiology* (3rd Ed.). Cambridge: Cambridge University Press.
- Cacioppo, J. T., Tassinary, L.G., & Fridlund, A. (1990). The skeletalmotor system. In In J. T. Cacioppo & L.G. Tassinary. *Principles of Psychophysiology* (Eds.) *Principles of Psychophysiology*. pp. 325--84. Cambridge University Press.
- Campbell, D. T., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.
- Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Avon Books.
- Darwin, C. (1965 [1872]). *The expression of the emotions in man and animals*. Chicago: University of Chicago Press.
- Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, 130, 355–391.

- Dickinson, A., & Dearing, M. F. (1979). Appetitive-aversive interactions and inhibitory processes. In A. Dickinson & R. A. Boakes (Eds.), *Mechanisms of learning and motivation* (pp. 287–324). New York: Academic Press.
- Dienstbier, R. A. (1989). Arousal and physiological toughness: Implications for mental and physical health. *Psychological Review*, 96, 84–100.
- Ekman, P. (1993). Facial expression of emotion. *American Psychologist*, 48, 384–392.
- Fazio, R. H. (1989). On the power and functionality of attitudes: The role of attitude accessibility. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 153–179). Hillsdale, NJ: Erlbaum.
- Fazio, R. H., Chen, J., McDonel, E. C., & Sherman, S. J. (1982). Attitude accessibility, attitude-behavior consistency, and the strength of the object-evaluation association. *Journal of Experimental Social Psychology*, 18, 339–357.
- Feldman-Barrett, L. (2006). Emotions as natural kinds? *Perspectives on Psychological Science*, 1, 28–58.
- Fridlund, A. J., Schwartz, G. E., & Fowler, S. C. (1984). Pattern recognition of self-reported emotional states from multiple-site facial EMG activity during affective imagery. *Psychophysiology*, 21, 622–637.
- Goffman, E. (1963). *Stigma: Notes on the management of spoiled identity*. New York: Simon & Schuster.
- Gray, J. A. (1982). *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system*. Oxford: Oxford University Press.
- Heinze, H.-J., Munte, T. F., & Mangun, G. R. (Eds.). (1994). *Cognitive electrophysiology*. Boston: Birkhauser.
- Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, 76, 359–364.
- Kasprowicz, A. L., Manuck, S. B., Malkoff, S. B., & Krantz, D. S. (1990). Individual differences in behaviorally evoked cardiovascular response: Temporal stability and hemodynamic patterning. *Psychophysiology*, 27, 605–619.
- Katz, D. (1960). The functional approach to the study of attitudes. *Public*

Opinion Quarterly, 24, 163–204.

Kelsey, R. M., & Guethlein, W. (1990). An evaluation of the ensemble averaged impedance cardiogram. *Psychophysiology*, 28, 24–33.

Konorski, J. (1967). *Integrative activity of the brain: An interdisciplinary approach*. Chicago: University of Chicago Press.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review*, 97, 377–395.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1992). A motivational analysis of emotion: Reflex-cortex connections. *Psychological Science*, 3, 44–49.

Lasswell, H. D. (1936). Certain changes during trial (psychoanalytic) interviews. *Psychoanalytic Review*, 23, 241–247.

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. New York: Springer.

Lerner, M. J. (1980). *The belief in a just world: A fundamental delusion*. New York: Plenum Press.

Lerner, M. J., & Miller, N. H. (1978). Just world research and the attribution process: Looking back and ahead. *Psychological Bulletin*, 85, 1030–1051.

Manuck, S. B., Kamarck, T. W., Kasprowicz, A. S., & Waldstein, S. R. (1993). Stability and patterning of behaviorally evoked cardiovascular reactivity. In J. Blascovich & E. S. Katkin (Eds.), *Cardiovascular reactivity to psychological stress and disease: An examination of the evidence* (pp. 83–108). Washington, DC: American Psychological Association.

Marshall-Goodell, B. S., Tassinary, L. G., & Cacioppo, J. T. (1990). Principles of bioelectric measurement. In J. T. Cacioppo & L. G. Tassinary (Eds.), *Principles of psychophysiology: Physical, social, and inferential elements* (pp. 113–148). New York: Cambridge University Press.

Masterson, F. A., & Crawford, M. (1982). The defense motivation system: A theory of avoidance behavior. *The Behavioral and Brain Sciences*, 5, 661–696.

McNaughton, N. (1993). Stress and behavioral inhibition. In S. C. Stanford & P. Salmon (Eds.), *Stress: An integrated approach* (pp. 91–109). New York: Academic Press.

- Mendes, W.B., Blascovich, J., Lickel, B., & Hunter, S. (2002). Challenge and threat during interactions with White and Black men. *Personality and Social Psychology Bulletin*, 28, 939–952.
- Mittleman, B., & Wolff, H. G. (1939). Affective states and skin temperature: Experimental study of subjects with “cold hands” and Raynaud's syndrome. *Psychosomatic Medicine*, 1, 271–292.
- Nielsen-Bohlman, L., & Knight, R. T. (1994). Event-related potentials dissociate immediate and delayed memory. In H. J. Heinze, T. F. Munte, & G. R. Mangun (Eds.), *Cognitive electrophysiology* (pp. 169–182). Boston: Birkhauser.
- Obrist, P. A. (1981). *Cardiovascular psychophysiology: A perspective*. New York: Plenum.
- Russell, J. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110, 145–172.
- Saxe, L., Dougherty, D., & Cross, T. (1987). Lie detection and polygraph testing. In L. S. Wrightsman, C. E. Willis, & S. Kassin (Eds.), *On the witness stand* (pp. 14–36). Newbury Park, CA: Sage.
- Schneirla, T. C. (1959). An evolutionary and developmental theory of biphasic processes underlying approach and withdrawal. In *Nebraska Symposium on Motivation: 1959* (pp. 1–42). Lincoln: University of Nebraska Press.
- Seery, M. D. (2011). Challenge or threat? Cardiovascular indexes of resilience and vulnerability to potential stress in humans. *Neuroscience and Biobehavioral Reviews*, 35, 1603–1610.
- Seery, M. D., Weisbuch, M., Hetenyi, M. A., & Blascovich, J. (2010). Cardiovascular measures independently predict performance in a university course. *Psychophysiology*, 47, 535–539.
- Shapiro, D., & Crider A. (1969). Psychophysiological approaches in social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 3, pp. 1–49). Reading, MA: Addison-Wesley.
- Shaver, P. R., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further explorations of a prototype approach. *Journal of Personality and Social Psychology*, 52, 1061–1086.

- Sherwood, A. (1993). Use of impedance cardiography in cardiovascular reactivity research. In J. Blascovich & E. S. Katkin (Eds.), *Cardiovascular reactivity to psychological stress and disease: An examination of the evidence* (pp. 157–200). Washington, DC: American Psychological Association.
- Sherwood, A., Allen, M. T., Fahrenberg, J., Kelsey, R. M., Lovallo, W. R., & van Doornen, L. J. P. (1990). Methodological guidelines for impedance cardiography. *Psychophysiology*, 27, 1–23.
- Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69, 797–811.
- Swann, W. B. (1983). Self-verification: Bringing social reality into harmony with the self. In J. Suls & A. G. Greenwald (Eds.), *Social psychological perspectives on the self* (Vol. 2, pp. 33–66). Hillsdale, NJ: Erlbaum.
- Tomaka, J., & Blascovich, J. (1994). Effects of justice beliefs on cognitive appraisal of and subjective, physiological, and behavioral responses to potential stress. *Journal of Personality and Social Psychology*, 67, 732–740.
- Tomaka, J., Blascovich, J., Kelsey, R. M., & Leitten, C. L. (1993). Subjective, physiological, and behavioral effects of threat and challenge appraisal. *Journal of Personality and Social Psychology*, 65, 248–260.
- Tomaka, J., Blascovich, J., Kibler, J., & Ernst, J. M. (1997). Cognitive and physiological antecedents of threat and challenge appraisal. *Journal of Personality and Social Psychology*, 73, 63–72.
- Troland, L. T. (1929). *The principles of psychophysiology: A survey of modern scientific psychology* (Vols. 1–3). New York: Van Nostrand.
- Vanman, E. J., Paul, B. Y., Ito, T. A., & Miller, N. (1997). The modern face of prejudice and structural features that moderate the effect of cooperation on affect. *Journal of Personality and Social Psychology*, 73, 941–959.
- Vanman, E. J., Saltz, J.L., Nathan, L.R., & Warren, J.A. (2004). Racial discrimination by low-prejudiced Whites' facial movements as implicit measures of attitudes related to behavior. *Psychological Science*, 15, 711–714.
- Vrana, S. R., & Lang, P. J. (1990). Fear imagery and the startleprobe reflex. *Journal of Abnormal Psychology*, 99, 181–189.

Vrana, S. R., Spence, E. L., & Lang, P. J. (1988). The startleprobe response: A new measure of emotion? *Journal of Abnormal Psychology*, 97, 487–491.

Waid, W. M. (1984). *Sociophysiology*. New York: Springer-Verlag.

* Correspondence should be addressed to Jim Blascovich, Department of Psychology, University of California, Santa Barbara, CA 93106.

¹ Jones and Sigall (1971) exploited this mystique very creatively. They convinced research participants that physiological measures recorded through electrode sensors connecting their bodies to the researchers' sophisticated-looking physiological recording devices would reveal their true thoughts and feelings even though the researchers did not really record physiological responses. Their "bogus pipeline" presumably motivated human participants to self-report what the machines would supposedly reveal objectively. The rationale underlying the bogus pipeline underlies the success of much lie detection work. Whether or not physiological measures index psychological constructs veridically, when individuals believe that they do, they are more likely to reveal their hidden thoughts and feelings. Most lie detection experts, or "polygraphers" as they prefer, realize this fact. Their success stems more from confessions of likely suspects than from physiological patterns unequivocally associated with truth and lying actually recorded from suspects (Saxe, Dougherty, & Cross, 1987). Jones and Sigall also realized this fact.

² Readers are urged to access other sources such as Blascovich & Mendes (2010); Blascovich, Vanman, Mendes, & Dickerson (2011); and Cacioppo, Tassinari & Berntson (2007).

Chapter seven Research Methods in Social and Affective Neuroscience

Elliot T. Berkman, William A. Cunningham and Matthew D. Lieberman

Introduction

In the two decades since the term “social neuroscience” was first used (Cacioppo & Berntson, 1992), social and personality psychologists have increasingly and enthusiastically adopted neuroscience methods including neuroimaging, endocrinology, and peripheral physiological measurement to address social psychological questions. The number of Google hits for the phrase “social neuroscience” rose from 393 in 2001 to 290,000 in 2011, and from 6 to 946,000 for the phrase “social cognitive neuroscience” across the same period. There are now two societies for social neuroscience, each with its own journal. As of 2012, more than 60 labs around the world identify themselves with social neuroscience, social cognitive neuroscience, or social-affective neuroscience, using neuroimaging to inform social psychological theories on such topics as self-knowledge, theory of mind, mentalizing, emotion regulation, empathy, and implicit attitudes (Lieberman, 2012).

Despite its increasing popularity, there remains a gap in the level of methodological knowledge between those who practice social neuroscience and those who comprise a large part of the intended audience but are not active researchers in the field, namely social-personality psychologists. This may be in part because the rapid and recent development of the field has outpaced changes in doctoral curricula and faculty hiring in psychology departments. In our experience, doctoral students, postdocs, and faculty alike in social and personality psychology are interested in social neuroscience and optimistic about its use to further psychological theory. They just may not have had the chance to learn about its methods in sufficient detail to be informed consumers or to integrate neuroscience into their own research.

Our aim here is to provide a concise and up-to-date description of neuroimaging methods used by social neuroscientists – particularly functional

magnetic resonance imaging (fMRI) and electroencephalography (EEG) – for an audience of social and personality psychologists. This chapter is not meant to be a comprehensive “how to” guide for practitioners, but rather a “what, how, and why” guide for consumers of social neuroscience research, particularly graduate students and faculty in psychology and related fields who are curious about but unfamiliar with the methods. For more comprehensive guides, we recommend *Methods in Social Neuroscience* (Harmon-Jones & Beer, 2009) for an overview of various methods, as well as *Functional Magnetic Resonance Imaging* (Huettel, Song, & McCarthy, 2009) and *Handbook of Functional MRI Data Analysis* (Poldrack, Mumford, & Nichols, 2011) for the details of fMRI acquisition and analysis, respectively. For an in depth discussion of EEG methods, we recommend *An Introduction to the Event-Related Potential Technique* (Luck, 2005). We focus on fMRI and EEG because they are the main neuroimaging modalities for social and affective neuroscience. Although we do not explicitly cover other modalities that are also commonly used, such as structural MRI (sMRI; Mechelli, Price, Friston, & Ashburner, 2005) and transcranial magnetic stimulation (TMS; Hallett, 2007), we note that many of the conceptual issues that we discuss with respect to fMRI and EEG are also applicable to these and other imaging modalities (e.g., positron emission tomography).

This chapter reviews two methods in the context of eight conceptual questions in social neuroscience. First, we briefly discuss what kinds of questions can be answered using social neuroscience methods. Next, we describe how fMRI studies are designed and how their data are collected, analyzed, and reported. A section on EEG and event-related potential (ERP) studies follows. Finally, we turn to a discussion of recent debates and controversies in social neuroscience and related fields. Each section is written to be independent of the others, so readers can skip sections according to their content interest or desired level of detail.

Types of Questions That Social Neuroscience Methods can Answer

Social-personality psychologists have been drawn to neuroimaging methods to answer at least three classes of questions relevant to social and personality theory (Lieberman, 2010). First, what are the neural mechanisms of social/personality processes? Second, do the neural systems involved in these

processes overlap with the neural systems of other processes in ways that might not be easily discovered with other methods? And third, are there neural systems that are uniquely involved in social/personality processes? We consider each briefly in the following sections.

Brain Mapping: What Are the Neural Mechanisms of Social-Personality Processes?

The ability to make inferences about which mental processes are associated with a given pattern of neural activation depends critically on the quality of our understanding of the function of various brain regions. The more detailed understanding we have of what a region does, the stronger our inference can be about which mental process is occurring when we observe activation in that region. Because of this fact, the enterprise of social neuroscience actually consists of two steps that happen in an iterative loop: using brain mapping to understand which brain regions are involved in a given psychological process, then testing psychological theory based on this new information and updating it when necessary (Cunningham, 2010). For example, if we are highly confident that the left ventrolateral prefrontal cortex and the medial temporal lobes are involved in processing depth during memory encoding, and observe activation in these regions during nonsocial information encoding but in a distinct set of brain regions during encoding of social information, then we can infer that social information is encoded in a qualitatively different way than nonsocial information (Mitchell, Macrae, & Banaji, 2004).

The key point is that high-quality brain mapping is critical for social neuroscience to be successful. Considerable progress has been made in this area in recent decades, but even more work remains. For instance, we know that the right inferior frontal gyrus is involved in self-control (Aron, Robins, & Poldrack, 2004), but it is also involved in related processes such as task switching (Cools, Clark, Owen, & Robbins, 2002), conditional rule use (Bunge & Zelazo, 2006), and competitive biasing of abstract information (Munakata, Herd, Chatham, Depue, Banich, & O'Reilly, 2011). These kinds of brain-mapping studies are and will continue to be essential in triangulating a more precise understanding of what computations this region is performing, and their results are directly informative to social psychological theory on self-control. In general, these refinements in brain-mapping knowledge include: identifying the relationship between mental process and function or connectivity of a region; specifying boundary conditions or contextual moderators; and charting

functional co-activations among two or more regions that are characteristic of a specific process. Each contributes to our understanding of the mapping between mind and brain, and thus each is a critical part of social neuroscience.

Convergences: How Do Social-Personality Neural Systems Overlap with Other Systems?

Identifying convergences in the brain systems of mental processes that subjectively feel distinct or are otherwise predicted to be different is another way that neuroscience can contribute to psychological theory. Examples of convergences include strong overlaps in the brain systems involved in social and nonsocial rewards (Izuma, Saito, & Sadato, 2008), and between social and physical pain (Eisenberger, 2012). This latter set of findings suggested a novel intervention for social pain – Tylenol administration – that would have been difficult to justify based on behavior and self-report data alone (DeWall et al., 2010). Convergences in neural function between different mental processes have also been used successfully to compare theories when each offers a competing process model. For example, psychologists have debated whether we understand the emotions of others by mirroring their experience directly or by first simulating their experience in our own mind and then projecting our emotions onto them (Wheatley, Milleville, & Martin, 2007). It turns out that these accounts – mirroring and self-projection – have distinct neural systems, and that only the system for self-projection is consistently active when subjects attempt to empathize with the emotions of another (Waytz & Mitchell, 2011). This overlap between the brain systems for experiencing an emotion directly and for understanding it in others has provided support for the self-projection theory of empathy.

We note again that the strength of inference based on neural convergence is limited by the specificity of brain mapping. A critical open question is: What level of precision on a neural level is required to infer that two processes are the same? Clearly, it is finer than the entire brain (i.e., “any two processes that both occur in the brain are the same” makes no sense), and probably finer than a hemisphere or a lobe, too. Are two processes that both activate the same 1 cm³ chunk of tissue the same? How about 1mm³? It is unlikely that two instances of even the *same* mental process share that level of specificity (see Points #4 and #5 on replication later in the chapter), but the exact threshold for ontological “sameness” is currently unknown, and may be beyond the resolution of our current neuroimaging technologies. This issue precludes definitive claims about

shared mechanism for now, but these will improve along with advances in methods and knowledge.

Divergences: Are There Brain Systems Unique to Social-Personality Processes?

Neuroscience methods can also be used to find surprising divergences in the mechanisms of various processes that would otherwise seem similar. An emerging theme in social neuroscience is that the brain networks for processing social stimuli may be distinct from those that process nonsocial stimuli. The study by Mitchell *et al.* (2004) cited earlier is an example: social and nonsocial memory encoding did not merely differ quantitatively on depth of processing, but also qualitatively in that they used distinct neural regions. Memory retrieval was correlated with one set of regions for nonsocial stimuli and with an entirely different set of regions for social stimuli.

We note that the social/nonsocial distinction is just one potential divergence uncovered thus far by social neuroscience. It happens to have considerable support, but there are others. One example comes from the emotion literature, where LeDoux (1996) famously identified two distinct pathways, the “low road” and the “high road,” to achieve a fear response. Another comes from social cognition, in which Foerde, Knowlton, and Poldrack (2006) demonstrated that implicit and explicit learning (differentiated with a cognitive load manipulation) produced equivalent behavioral results but used distinct neural regions to do so. Social neuroscience is still young; surely more surprising divergences are yet to be discovered.

fMRI Methods

Study Design

In addition to the usual considerations that one must take into account when designing a social-personality psychology study, there are several special design considerations that are unique to fMRI. These can be classified into three categories that we call statistical, technological, and human factors.

The main statistical considerations derive from the facts that all fMRI studies contain at least one within-subjects factor (because, as discussed later in the chapter, the raw data are in arbitrary units that vary considerably from subject to

subject), and that the data are arranged in a time-series with a potentially strong autocorrelation (i.e., later time points cannot be assumed to be independent of earlier time points). Another factor is that the variable that is measured by fMRI, the blood oxygenation level-dependent (BOLD) signal, is slowed and elongated in time relative to the stimuli used in our experiments. For example, even a very brief (e.g., 100ms) burst of white noise will generate a BOLD signal in auditory cortex that begins 3–5 seconds following onset and peaks around 6–8 seconds following onset, with a total duration of approximately 6–8 seconds (Glover, 1999; see Figure 7.1). Longer stimuli will produce correspondingly longer BOLD responses, and several stimuli presented in rapid succession will produce overlapping BOLD responses that together account roughly to one cumulative BOLD response (approximating the rules of a linear system within a certain range).

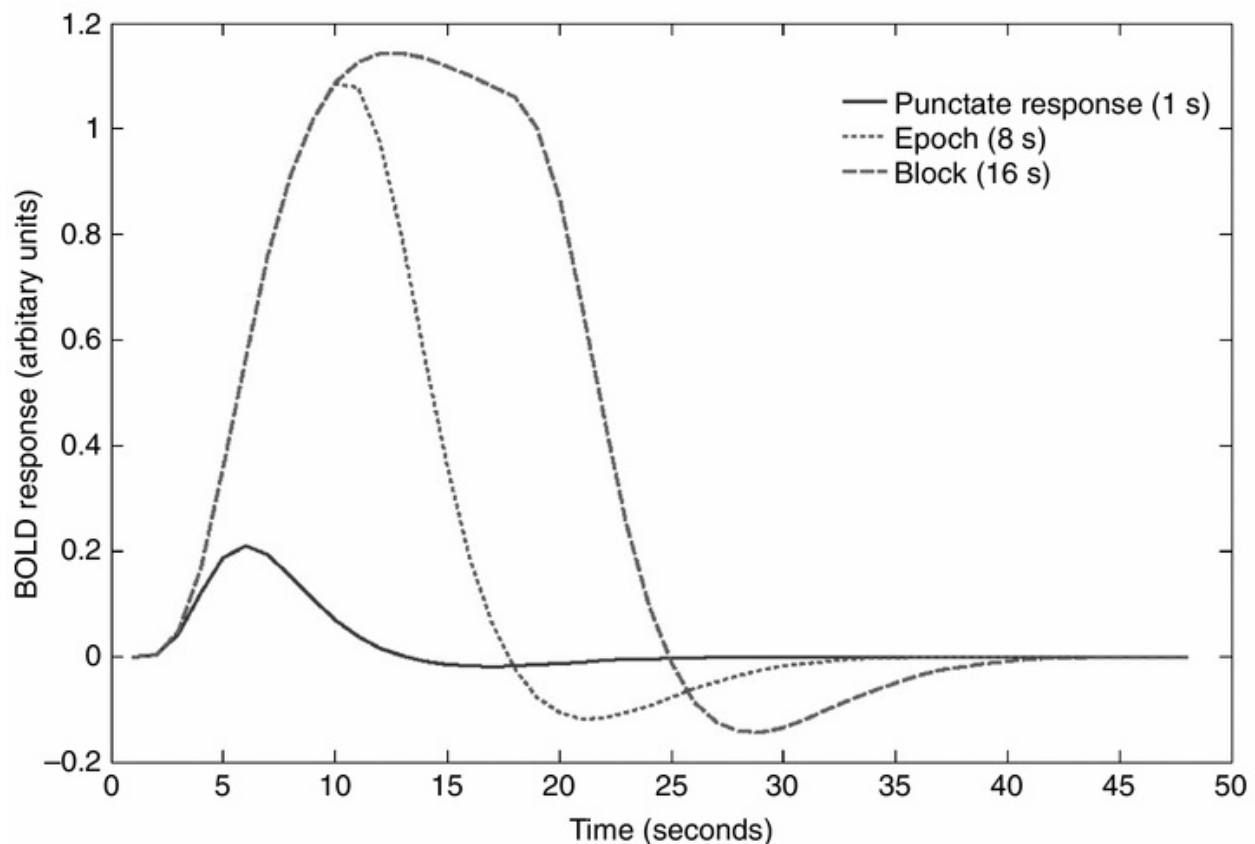


Figure 7.1. The predicted BOLD response for three stimuli of varying duration. The response for a very brief stimulus of 1 second or shorter, called a stick function or punctate event (solid line). The response for a stimulus in the 6–8 second range, called an epoch (dotted line). And the response for a stimulus of 16 seconds or longer, called a block (dashed line).

Together, these facts imply that stimuli presented too close to one another in time or in regular succession will be confounded in the BOLD response and difficult statistically to differentiate from one another. There are two main design solutions to this problem: blocked and event-related designs (Figure 7.2). In a blocked design, trials of the same type are grouped together into relatively long “blocks” that are usually 20 to 50 seconds each, and separated by a fixation cross “resting” baseline, various other conditions, or both. Blocked designs maximize the efficiency to detect differences between conditions throughout the brain, but provide no information about the shape of the curve of the BOLD response over time, called the hemodynamic response curve (Buckner, Bandettini, O’Craven, Savoy, Petersen, Raichle, & Rosen, 1996). The other solution is to use a rapid event-related design in which brief events (typically 1–5 seconds each) are presented in close succession. (A third type of design, slow event-related, which can be thought of as a hybrid of blocked and rapid event-related designs, has fallen out of favor because of its low detection power.) Rapid event-related designs are popular because, when used appropriately, they allow for a reasonable balance between signal detection and hemodynamic response estimation efficiency (Birn, Cox, & Bandettini, 2002), and allow stimuli in various conditions to intermingle in ways that may be advantageous psychologically (e.g., in studies on emotion where habituation to negatively valenced stimuli can be a problem). Fortunately, researchers can either use “jitter” (insert a systematically varying amount of time between successive stimuli of the same trial type; Dale & Buckner, 1997) or optimize the order of the trials (Wager & Nichols, 2003) to both remove the temporal correlation between conditions and also tease apart the BOLD signal corresponding to each. In other words, by varying the order of trials or the length of the gap between trials, or even between parts of trials, statistical procedures are better able to recover the contrast between particular trials (or trial components).

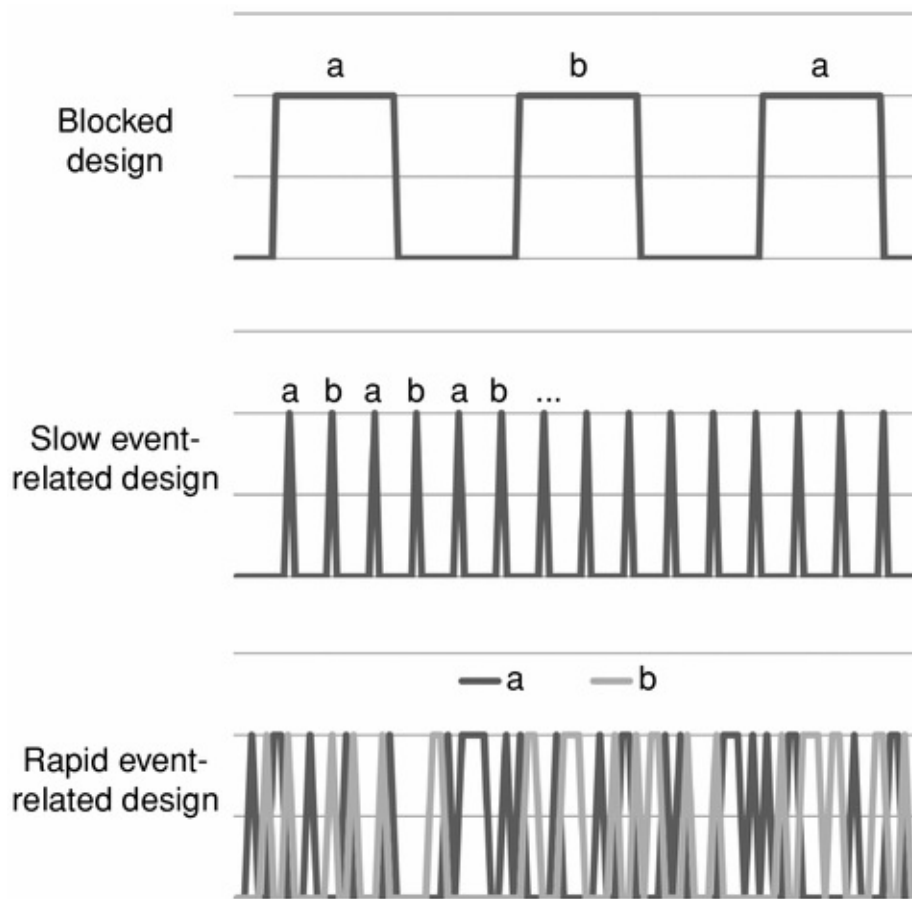


Figure 7.2. The main types of designs used in social neuroscience studies. (Top) Blocked designs feature the same kind of trial for prolonged periods (e.g., 30–50 seconds) followed by alternating periods of baseline. Blocked designs maximize power to detect differences among conditions but provide no information about the shape of the hemodynamic response function. (Middle) Slow event-related designs feature brief epochs of a single even followed by extended rest to allow the BOLD response to return to baseline. These designs are not often used because of their low power to detect differences between conditions (but they provide excellent estimation of the hemodynamic response function). (Bottom) Rapid event-related designs feature multiple trial types interleaved in a random or optimized sequence along with a baseline condition. These kinds of designs are popular because they offer a balance between contrast detection power and estimation power.

One last important statistical consideration is power. What sample size is sufficient and how many trials per condition are required? These questions are difficult to answer analytically because the variance of fMRI data and the effect size ranges widely from study to study and from brain region to brain region.

However, researchers are beginning to develop methods to conduct power analyses based on formal statistical models (Mumford & Nichols, 2008) and data simulations (Desmond & Glover, 2002; Murphy & Garavan, 2004). We anticipate especially rapid accumulation of knowledge in this area in the coming years because of the high demand relative to the current information available. For the time being, we are reluctant to provide informal rules of thumb because, just as with behavioral studies, the sample size required to achieve a given amount of power can vary considerably from study to study depending on the amount of noise, and the covariance between within-subjects conditions, among other factors. We provide further information about sample size in the section on between-subjects correlations (Point #2 later in the chapter).

The technological design factors concern the limitations of the fMRI machine and its environment. First, fMRI scanning is expensive: an hour of scanner time typically costs between \$400 and \$800 in the United States. In that time, a researcher must leave 10–15 minutes for subject setup inside the scanner, 10 minutes for various structural scans and other scans that are not directly relevant to the task, and several minutes for instructions. Efficient researchers can maximize the amount of scanner time spent on experimental tasks, but even in this case the tasks are shorter than they otherwise would be (e.g., in a behavioral laboratory study). This can present a problem for behavioral experimental paradigms that didn't originate in neuroscience and that require a large number of trials and/or subjects (either of which increases the total cost of an experiment). Generally speaking, it is more cost-effective to run fewer subjects each with more trials (vs. more subjects each with fewer trials), so social neuroscientists rely almost exclusively on within-subjects designs – even for experimental paradigms that are usually between-subjects – and try to employ manipulations that maximize the difference between conditions with as few trials as possible.

Second, fMRI scanning is loud. Tasks that rely on auditory cues (e.g., the auditory stop-signal task) must be adjusted accordingly and tested to ensure that task-relevant cues are not presented in the same frequency range as the scanner noise that accompanies functional MRI. And third, fMRI scanning is awkward. The most common type of study involves a single participant, alone and supine in the bore of the scanner, wearing goggles or a mirror for visual stimuli and headphones for auditory stimuli, holding a customized response device such as a scanner-safe button box, mouse, or joystick, and attempting to hold as still as possible. Because MRI relies on a very strong magnetic field (usually 1.5 to 4 Tesla), no ferromagnetic objects (e.g., most computers, cameras, or input devices

typically used in behavioral studies) can be near the scanner. Some experiments can be run within these constraints, but many cannot. However, neuroscientists have recently made great strides in overcoming some of the limitations of this environment, for example by having another person in the scanner room to study physical contact (Coan, Schaefer, & Davidson, 2006), using eye-tracking to simulate gaze-contingent social interaction (Wilms, Schilbach, Pfeiffer, Bente, Fink, & Vogeley, 2010), and studying gustatory responses to food with a milkshake pump system (Stice, Yokum, Burger, Epstein, & Small, 2011) or an elaborate olfactory cue delivery device (Small, Gerber, Mak, & Hummel, 2005).

Finally, to add to the technological limitations listed previously, researchers seeking to use fMRI for social neuroscience must put human beings into the scanner environment. Many humans are uncomfortable in dark and narrow tubes, including young children and individuals with ADHD or claustrophobia, making those challenging populations to study (Pierce, 2011; Redcay, Kennedy, & Courchesne, 2007) and precluding true random sampling of individuals. Further, from an embodied cognition perspective, lying motionless and flat on one's back may dampen some of the psychological processes that are of greatest interest to social and personality psychologists, such as approach motivation (Harmon-Jones & Peterson, 2009), and inhibiting head motion (which is typically desired to be at 2mm or less while inside the scanner bore) may lead to reductions in emotion experience (Davis, Senghas, Brandt, & Ochsner, 2010). Researchers could pretest designs in a realistic way (e.g., using a mock scanner setup) to ensure their manipulations are successful.

Data Acquisition

Participants are situated in the scanner in a supine position. Stimuli are presented via special scanner-safe goggles or a rear-projection system and headphones. Many scanning centers require participants to wear earplugs to dampen the noise, even if they are also wearing headphones. Depending on the task, participants may also hold a response device in one or both hands such as a button box or joystick. Scanning centers frequently take a variety of precautions to reduce head motion during the scan, ranging from foam padding and inflatable cushions to bite bars and custom face masks that strap to the scanner bed. During the scan, the experimenter remains immediately outside of the scanning room in a control room and can communicate with the participant through a microphone connected to the participant's headphones. Direct auditory communication with participants is typically limited during the scan because of

scanner noise, so participants are given an emergency call exit button in case they need to discontinue the scan for any reason.

Once situated in the scanner, participants' brain activation during the experimental tasks is measured indirectly by the fMRI scanner. Neural activity increases local blood flow, leading to increases in the oxygenated-to-deoxygenated hemoglobin ratio that perturbs the local magnetic field, and it is these small fluctuations in the local magnetic field that are detected in fMRI. Although the BOLD signal is an indirect measure of neural activity, there is now considerable evidence using a combination of optical imaging (Kennerley, Berwick, Martindale, Johnston, Papadakis, & Mayhew, 2005) and direct electrical recording (Viswanathan & Freeman, 2007) that the BOLD response is tightly coupled to synaptic activity and particularly to electrical local field potentials (Goense & Logothetis, 2008).

Even though there are now dozens of ways to use MRI scanners to examine changes in blood flow and volume, nearly every social neuroscience study employs wholebrain echo-planar imaging, which refers to the particular sequence of spatially varying magnetic fields and radio frequency pulses used by the fMRI scanner to collect the data. The considerable advantage of this technique is that it can be used to acquire an entire two-dimensional brain image in as little as 50 milliseconds, enabling wholebrain scanning (i.e., three-dimensional images) in less than 2 seconds. Each two-dimensional image, also known as a *slice*, typically has a resolution of 64×64 , containing a total of 4,096 volume elements, or *voxels*. As shown in Figure 7.3, wholebrain images are comprised of a series of slices that are acquired either sequentially (i.e., from bottom up or from top down) or interleaved (i.e., slices 1, 3, 5, etc., then slices 2, 4, 6, etc) and then stitched together to form a single three-dimensional image (or *volume*). A typical volume contains 30–35 slices, totaling 120,000 to 150,000 voxels. In the language of MRI, the time between subsequent volumes is known as the repetition time, or TR. Ultimately, a raw fMRI data set consists of a time-series of wholebrain images acquired at each TR (usually once every 2 seconds or so) for the duration of the task (usually 5–10 minutes, or 150–300 scans).

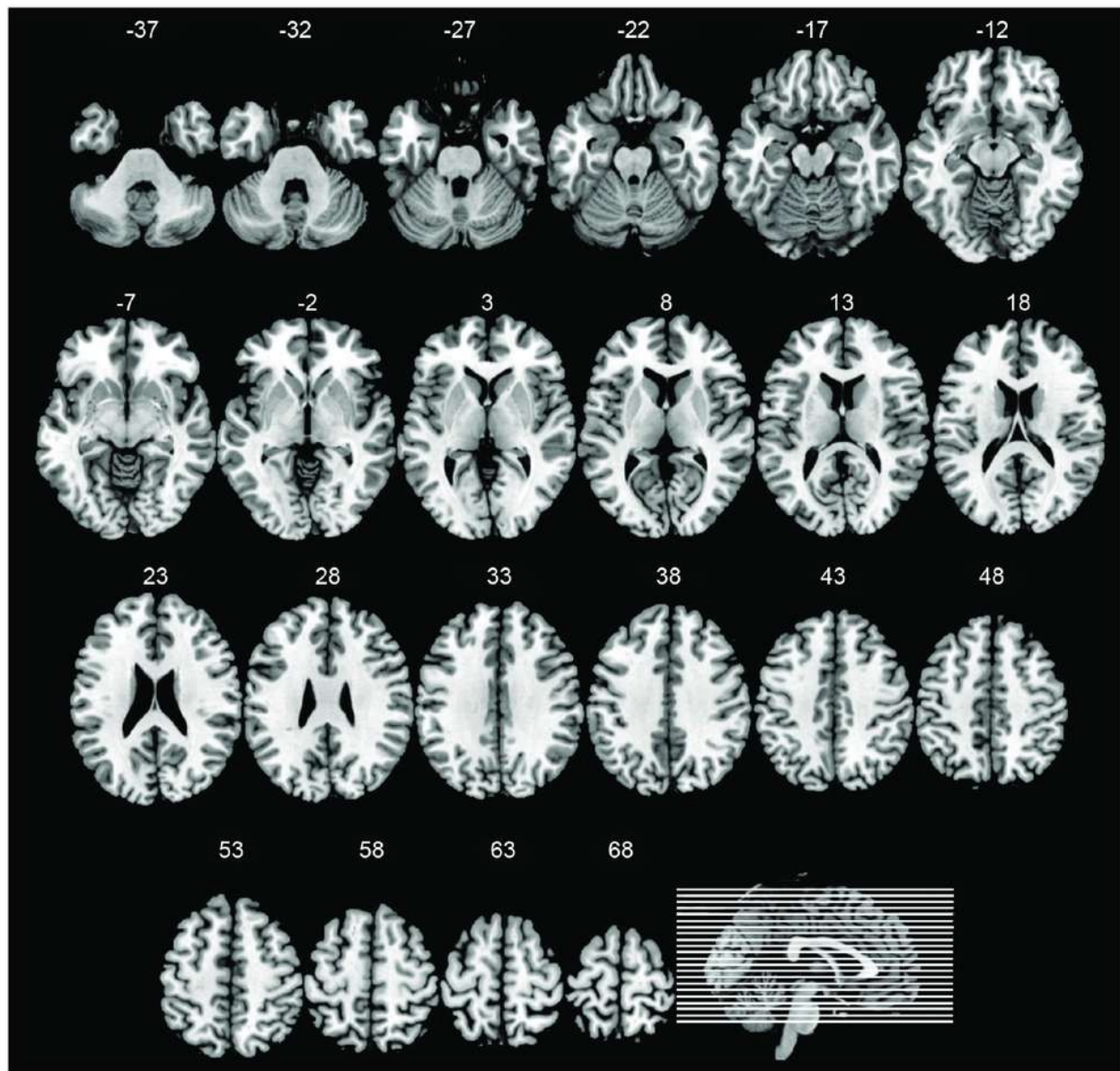


Figure 7.3. Wholebrain images are constructed from a series of two-dimensional “slices.” The figure shows 22 axial (horizontal) slices through a template brain in Montreal Neurological Institute (MNI) coordinates. The slices are arranged from inferior (top left) to superior (bottom), with the sagittal (side view) image on the bottom right indicating the position of each slice.

The gain of such rapid scanning comes at a cost. Echo-planar imaging is extremely sensitive to distortions, or inhomogeneities, in the magnetic field. These distortions are particularly strong at boundaries between tissue and air-filled cavities, and cause susceptibility artifacts in the images – regions where

the signal is poor or unusable. Unfortunately, some of the regions that are of greatest interest to social neuroscientists such as orbitofrontal cortex (OFC; Beer, John, Scabini, & Knight, 2006), the temporal poles, and the medial temporal lobe including the amygdala (Olson, Plotzker, & Ezzyat, 2007) are prone to susceptibility artifacts because of their proximity to the nasal cavity and the auditory canal, respectively. Fortunately, several methods have been developed recently to minimize signal loss in these regions (e.g., using an oblique axial-coronal acquisition angle; Weiskopf, Hutton, Josephs, Turner, & Deichmann, 2007), and are increasingly adopted in social neuroscience (e.g., Hare, Camerer, & Rangel, 2009; Xu, Monterosso, Kober, Balodis, & Potenza, 2011).

Functional imaging data are nearly always collected in tandem with at least one structural image that has relatively higher resolution (e.g., voxel dimensions of 1mm^3 for structural vs. 3mm^3 for functional images). These images serve as a reference map to help localize activations identified using the functional images and for presentation purposes, which is why they are often described as “anatomical” images. For example, the “blobs” of activation found using the EPI images are often displayed overlaid on a group average structural image (see Figure 7.4a), which helps readers identify the brain regions involved. Using a group average overlay for presentation is preferred to using a template (Figure 7.4b), which is often more aesthetically pleasing but misleading regarding the true resolution obtained in the study. There are dozens if not hundreds of varieties of structural images available, but the magnetization-prepared rapid gradient echo, or MP-RAGE, is most commonly used in social neuroscience studies because of its high resolution, excellent gray-to-white matter contrast, and fast acquisition time (Brant-Zawadzki, Gillan, & Nitz, 1992).

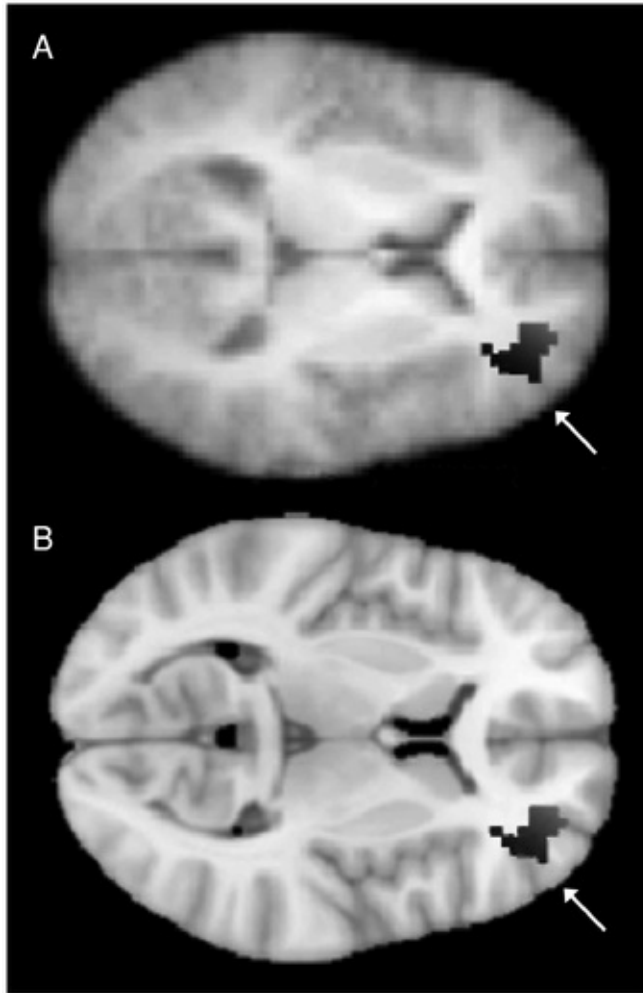


Figure 7.4. (a) An activation “blob” overlaid on a group mean anatomical image (an MP-RAGE) provides good spatial localization and is preferred over using (b) a template overlay, which can be misleading about the resolution of the acquired data.

Data Cleaning and Preprocessing

Following acquisition, fMRI data must go through a number of cleaning, shuffling, and normalizing steps collectively known as *preprocessing* before they are ready for analysis (Figure 7.5). Preprocessing is computationally elaborate, memory intensive, and still under active mathematical and theoretical development. Because of these challenges, the number of software packages available to preprocess and analyze fMRI data is ever increasing. The most common are Statistical Parametric Mapping (SPM), Functional MRI of the Brain Centre Software Library (FSL), Analysis of Functional NeuroImages (AFNI), and FreeSurfer, which are all free, and BrainVoyager, which is proprietary. Most

labs use SPM, FSL, or a combination of those two (see Poldrack, Mumford, & Nichols, 2011, for a comparison). There are many others. To complicate matters further, the consensus within the research community about how best to preprocess and analyze data is still in flux, and each of these packages is constantly updated to reflect these changing opinions. Also, each software package preprocesses raw data in a slightly different order. Thus, instead of describing any one piece of software or processing algorithm in detail, we focus on the basic steps of analysis that are implemented in most or all of the packages (though not necessarily in the order described here). The specific software and computational methods will surely change in the coming years, so our aim is to explain the purpose of each step more than to describe how that step is calculated.

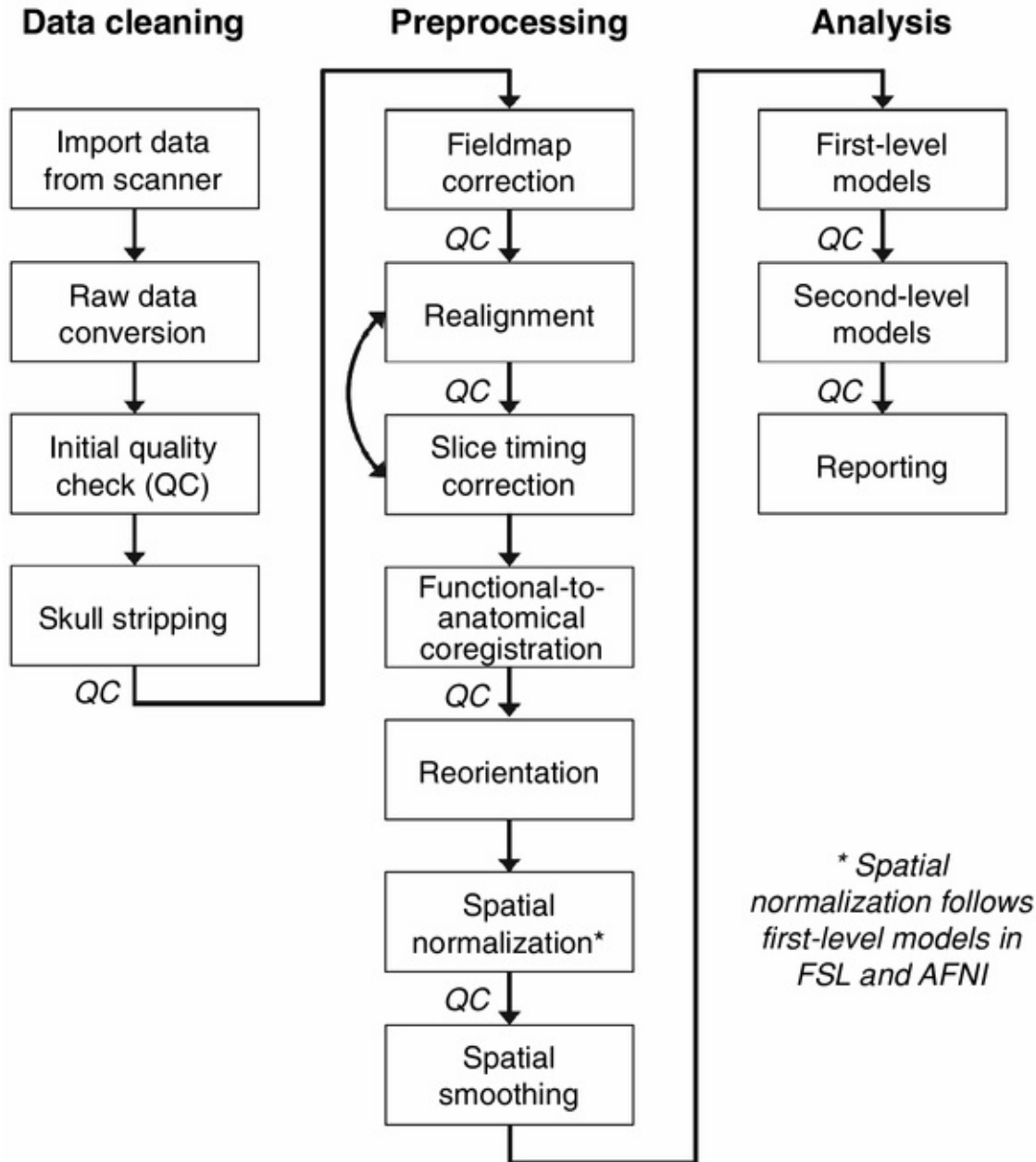


Figure 7.5. Data flow of a typical social neuroscience fMRI study, including cleaning, preprocessing, analysis, and reporting steps. Note that the order of motion correction and slice timing correction are often swapped, and that SPM computes normalization before smoothing and first-level statistical analysis, whereas FSL and AFNI normalize after those two steps. QC = (additional) quality checks.

Raw data conversion. Data arrive off the scanner in a raw file format known as DICOM. There is one DICOM file for each wholebrain volume and several

for the anatomical scan. Each software package contains its own tool for converting the DICOMS to a more usable format, and there are several excellent stand-alone utilities freely available (e.g., MRIConvert). Most packages now use the a format developed by the Neuroinformatics Technology Initiative (NifTI) that allows for a time-series of wholebrain volumes from a task to be grouped into a single four-dimensional file (i.e., three spatial dimensions plus sequences of volumes over time).

After conversion, it is prudent to make a quality check of some kind, either by scanning through manually or using an automated tool. The main sources of contamination at this stage are participant motion and data spikes from ambient electromagnetic noise. Both of these may be correctable, but doing so requires extra steps beyond the standard preprocessing stream. This is important to note, as many of the tools allow start-to-finish batching of an entire preprocessing and analysis stream without forcing any data quality inspection. Without visual inspection, it is easy for noisy or distorted raw data to sneak by unexamined.

Field map correction. It is now common to begin preprocessing by correcting for distortions in the magnetic field. This can be done using a set of *field map* scans that contain a vector field of local magnetic inhomogeneities at each voxel. As noted earlier, attempting to correct for these inhomogeneities is particularly important if the hypotheses involve brain regions highly susceptible to dropout, such as the OFC. Even though the correction is not perfect, it can restore a considerable amount of signal, particularly if participant motion is low and a separate field map is obtained immediately before or after the functional run of interest (Hutton, Bork, Josephs, Deichmann, Ashburner, & Turner, [2002](#)).

Realignment. Next, researchers typically correct for motion between functional volumes and changes in timing within them. These steps are known as *realignment* and *slice-timing correction*, respectively. Realignment uses a six-parameter rigid body transformation (i.e., motion along the x, y, and z axes and rotation of pitch, roll, and yaw without warping or stretching) to quantify and subsequently correct for motion from volume to volume, similar to stacking a deck of cards into a neat pile after shuffling them. Slice-timing correction adjusts the slices within each volume to account for changes in signal that may have occurred within the TR, and is generally only necessary in studies where precise timing is critical such as event-related designs or those measuring functional connectivity. There is some debate regarding the best order to compute realignment and slice-timing correction as they are mutually dependent. For this reason, the field is moving toward consensus that an algorithm that computes

them simultaneously is needed (Smith, Jenkinson, Woolrich, Beckmann, Behrens, Johansen-Berg et al., 2004).

Coregistration. The purpose of the next step, *coregistration*, is to move the anatomical and/or the (now realigned) functional scans in space so they are perfectly overlapping with one another. (Note that the term “registration” is sometimes used to refer collectively to what we call coregistration and realignment. However, we will maintain a distinction between them throughout.) If realignment is akin to stacking a deck of cards, then coregistration is like placing the cut card on top; the cut card is qualitatively different than the other cards in the deck, but it has the same dimensions and must be oriented in exactly the same way as the other cards to be of any use. The algorithm used for coregistration is the same as the one used for realignment – the purpose is still to correct for motion – but coregistration adjusts for movement that occurs between the functional and structural scans. Considering this, it makes sense to compute coregistration separately from realignment because there can be a considerable amount of time between the functional and structural scans, depending on the experimental protocol. For example, the task of interest might take place at the very beginning of the scanning session, followed by several other tasks lasting 10–15 minutes each, with the anatomical scan at the end. Coregistration must account for the cumulative motion within each scan and also between scans (when participants are often encouraged to move in order to prevent within-task movement). Thus, acquiring the functional and anatomical scans in close proximity or minimizing between-task movement can facilitate high-quality coregistration.

Reorientation. Realignment and coregistration aim to put all of the acquired scans into the same space as one another; the next step, *reorientation*, is the first step toward placing the scans into a common space that can be used to generalize results and communicate them to other scientists. After conversion from DICOM files, the coordinate structure of the files is arbitrary. After successful realignment and coregistration, a given x,y,z coordinate on one image within the experiment will correspond to the same brain location on all the other images within the experiment but will not necessarily be the same as any other data set. And after reorientation, the brain images will all be angled and shifted in the same, standardized way. Specifically, the origin of the coordinate space (i.e., $x = 0$, $y = 0$, $z = 0$) will have been moved to a piece of white matter called the anterior commissure, the x - y plane will be parallel to the anterior commissure–posterior commissure (AC-PC) line, the y - z plane will delineate the interhemispheric fissure, and the x - z plane will be perpendicular to the other two

at the AC (Figure 7.6). For now, reorientation must be done manually because it involves mapping points in space to specific anatomical structures, and thus is highly labor intensive. Nonetheless, it is worth the effort because it can greatly improve normalization (the next step), which assumes that the images have been reoriented to the AC-PC line with the origin at the AC.

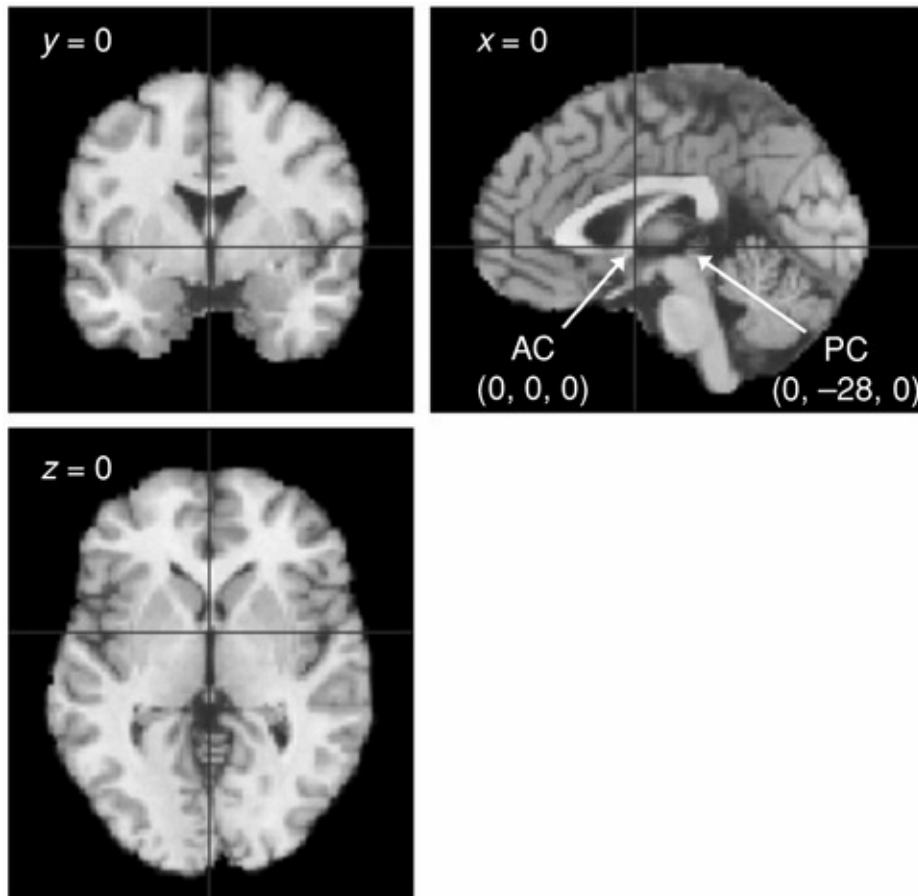


Figure 7.6. Three planes corresponding to the x , y , and z axes. (Top left) The coronal plane at $y = 0$. (Top right) The sagittal plan at $x = 0$. The anterior and posterior commissure (AC and PC, respectively) are visible along the horizontal crosshair. (Bottom left) The axial plane at $z = 0$. The crosshairs on all three axes are centered on the AC.

Normalization. Normalization adjusts for the simple fact that different people's brains are shaped differently. Group inference requires a way of averaging across people that preserves specificity of neural regions. For example, even though one subject's anterior cingulate might not be located at the same coordinates as that of another subject, we still want to average the activation in the anterior cingulate across those subjects without mixing in

nearby regions. One solution to this problem is to specify a priori a canonical brain with a known mapping between each point in space and anatomical landmarks (e.g., everyone agrees that the coordinates [18, -4, -20] on this brain correspond to the right amygdala). Then, find a way to map each participant's brain to this canonical brain. Once such a map is found, it becomes possible to generate a new version of the participant's brain that is in the canonical space. *Normalization* refers to this process of mapping then redrawing a brain into a common space.

Normalization is typically done into either space with a Talairach atlas (Talairach & Tournoux, 1988), or, more commonly, into space with a Montreal Neurological Institute atlas (MNI; Evans, Kamber, Collins, & MacDonald, 1994). The MNI atlas is generally preferred because it provides a probability atlas based on 152 brains (versus just one brain described in detail in the Talairach atlas), and there are several versions available for developmental populations (e.g., Fonov, Evans, Botteron, Almli, McKinstry, Collins, & the Brain Development Cooperative Group, 2011). Other advantages are that most software packages contain several high-resolution template brain images in MNI space, and there are an increasing number of databases available for labeling brain structures based on MNI coordinates (e.g., Fischl, Salat, Busa, Albert, Dieterich, Haselgrove et al., 2002; Shattuck, Mirza, Adisetiyo, Hojatkashani, Salamon, Narr et al., 2008).

Spatial smoothing. The last step of preprocessing is *spatial smoothing*, or blurring the brain images based on preset radius around each voxel. Smoothing accomplishes several things, including reducing outlier voxels (because they are blurred with nearby voxels), accounting for small differences in normalization between subjects, improving detection of activations larger than the smoothing kernel, and introducing a spatial correlation into the image (which is a prerequisite for some statistical methods). Smoothing is conceptually akin to averaging several related items on a personality questionnaire to get a better estimate of the true effect because the errors cancel each other out. The radius and the intensity of the smoothing is determined by a “smoothing kernel,” which typically follows a Gaussian distribution with a full width at half maximum (FWHM, or the width of the curve at half of its maximum value) of 4–8mm. It is recommended to make the smoothing kernel no larger than the smallest hypothesized activation, as larger kernels will decrease power to detect activations smaller than the kernel.

Data Analysis

Fully preprocessed images are entered into statistical analysis first on a subject-by-subject basis (*first-level* models) and then as a group for population inference (*second-level* models). This approach can be thought of as a proto-multilevel model where two levels (within-and between-subjects) are estimated separately (Schoemann, Rhemtulla, & Little, Chapter 21 in this volume). In this section we describe each of the two levels, the standard analyses that are computed within this framework, and some alternative models that have been adopted to address specific research questions.

First-level models. The subject-level models capture the entire time course of the task by breaking it down into component events (i.e., trials, rest periods, etc.). For example, the blocked design at the top of [Figure 7.2](#) could be described as: condition “A” beginning at seconds 10 and 130 and lasting 30 seconds each, condition “B” beginning at second 70 and lasting 30 seconds, and fixation baseline all other times. Even more complicated designs (e.g., the rapid event-related design at the bottom of [Figure 7.2](#)) can be described using lists of onsets and durations for each condition (called vectors in matrix algebra). In most software packages, any period of time that is not explicitly modeled is automatically included in a so-called implicit baseline condition, which is often used as a low-level comparison condition. For this reason, it is important (and sometimes forgotten!) to explicitly model every non-baseline event, even those of no interest, such as instruction periods. The on-off time course of each condition is then convolved with a canonical hemodynamic response function to create a predicted BOLD time course of a voxel that responds to only that condition, and those time courses are entered as regressors in a multiple regression model predicting the observed BOLD response ([Figure 7.7](#)). First-level models frequently also include several *covariates of no interest* that control for potentially confounding factors such as subject motion or peripheral physiological measures.

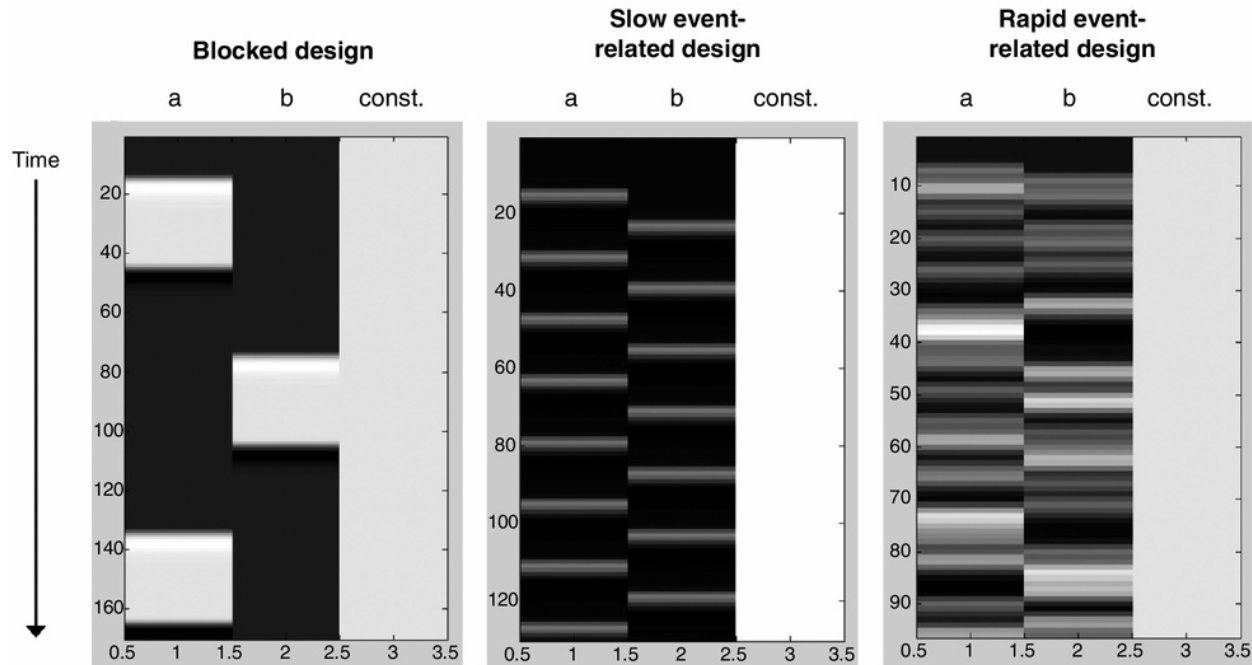


Figure 7.7. Statistical models for the three designs shown in Figure 7.2 (known as design matrices). (Left) Blocked design with two blocks of condition “a” and one block of condition “b.” (Middle) Slow event-related design with alternating epochs of conditions “a” and “b.” (Right) Rapid event-related design with a random ordering of “a,” “b,” and null events. By convention, time is shown from top to bottom. The rightmost column in each design matrix is the constant, and corresponds to β_0 in the regression model.

The solution to this regression model, which includes predictors for each condition, run, and covariate, is then estimated accounting for the autocorrelated nature of the errors, producing one beta weight for each condition. This approach is known as a “massively univariate general linear model” (Bowman, Guo, & Derado, 2007; Monti, 2011) because a regression model is created and solved at each voxel independently. The resulting beta weights (and corresponding t -values) are stored in three-dimensional maps representing the association between the predicted and observed BOLD response at each point in the brain.

Task effects are assessed via “contrasts” between conditions. This subtractive approach is necessary for neuroimaging based on the BOLD response because the raw data values are on an arbitrary scale, and thus only interpretable relative to one another. Contrasts in fMRI are exactly like those used in the context of ANOVA by social and personality psychologists in that they are weighted combinations of means from various conditions, only with beta values as the

dependent measure in each condition. For example, in a simple experiment with two conditions, A and B, if the beta weight at one voxel were 0.8 for A and 0.3 for B, the parameter estimate of the contrast, $[1 \ -1]$, which compares the model fit between the two conditions in that voxel, would be 0.5. A pairwise *t*-test can be used in this way to test for reliable differences between conditions. The vast majority of social neuroscience studies that examine the main effect of task conditions (or other comparisons that are based on within-subjects ANOVA such as interactions or custom linear combinations of conditions) use this *t*-test approach. For example, this strategy has been used to identify the neural systems involved in interpreting a series of cartoon panels in terms of mental state inference or physical motion (Brunet, Sarfati, Hardy-Bayle, & Decety, 2000).

Conjunction analysis is based on *t*-tests between conditions, as it involves examining the overlap between two or more pairwise *t*-tests. In a study with four conditions (e.g., a classic 2×2), conjunction identifies regions that are significantly more active during condition A versus B *and also* significantly more active during condition C versus D. This complements interaction analysis, which asks whether a region is differentially active during condition A versus B *to a greater extent than* during condition C versus D. Thus, conjunction is often used together with interaction in the same set of analyses. For example, McRae and colleagues used both conjunction and interaction analyses to identify the overlapping and distinct neural systems involved in two distinct forms of emotion regulation, distraction and reappraisal (McRae, Hughes, Chopra, Gabrieli, Gross, & Ochsner, 2010). First, distraction and reappraisal were each contrasted in a pairwise *t*-test to a “view negative” condition, generating Distraction versus View and Reappraisal versus View contrasts. A conjunction analysis between these two contrasts identified brain regions that were involved in both forms of emotion regulation, and an interaction analysis, (Distraction vs. View) versus (Reappraisal vs. View), identified brain regions that were differentially involved in one form greater than the other.

Another common analytic strategy is known as *parametric modulation* and is used to compare varying degrees of treatment effect within one condition. This is qualitatively different than the *t*-test approaches, because instead of comparing two or more conditions to each other, parametric modulation weighs each trial *within a single condition* according to some parameter (e.g., reaction time or Likert rating) – hence the name – and compares trials that are higher on that parameter to trials that are lower on that parameter. This analysis is independent from between-condition contrasts, and can be used to complement them or to replace them entirely depending on the design. A number of studies on reward

anticipation and responsivity use this approach to identify regions that are sensitive to increasing amounts of reward (e.g., where \$3 > \$2 > \$1, etc.) in addition to contrasting reward to loss (Eisenberger, Berkman, Inagaki, Rameson, Mashal, & Irwin, 2010; Knutson, Adams, Fong, & Hommer, 2001; Tom, Fox, Trepel, & Poldrack, 2007).

The last major class of within-subject models is used to pinpoint connectivity between regions, either at rest or as a function of condition. These models are alternately called *functional connectivity* or *effective connectivity*, and there are some subtle differences between them (Friston, 2009) that are not relevant for the present chapter. In resting-state connectivity, the observed BOLD response at rest from a “seed” region is used as the predictor in each subject's first-level design file (instead of a set of condition predictors). The resulting beta map estimates the strength of the linear relationship between the seed region and all other regions in the brain. This approach has been instrumental in identifying the so-called *default mode network* of regions that is preferentially active at rest in humans (Uddin, Kelly, Biswal, Castellanos, & Milham, 2009). In functional (or effective) connectivity, the BOLD response from a seed region is combined with task condition predictors to create a seed-by-condition interaction term (i.e., a psycho-physiological interaction, or PPI; Friston, Buechel, Fink, Morris, Rolls, & Dolan, 1997). This interaction term is the predicted BOLD response of a region that is more correlated with the seed region during some conditions than others. PPI is ideal to assess functional connectivity because it controls for resting state and anatomical connectivity to identify only regions whose correlation with the seed region changes as a function of task. This kind of analysis has been used in social neuroscience to find regions that are inversely related to emotion response areas (e.g., the amygdala) during emotion regulation but not other conditions (e.g., Berkman, Burklund, & Lieberman, 2009).

Second-level models. Once first-level models have been specified and estimated and contrasts have been generated for each subject in the sample, summary statistics for each subject (e.g., mean contrast values or mean connectivity estimates) are brought to a separate, group-level model for a random-effects analysis. These summary statistics often take the form of a wholebrain map of contrast values at each voxel (i.e., one map per subject per contrast), but can also be subsets of the wholebrain map known as *regions-of-interest*, or ROIs. Besides representing more spatially targeted hypotheses, the main advantage of analyzing ROIs at the second level is that they require fewer statistical comparisons compared to a wholebrain map. (See the section on statistical thresholding later in the chapter).

The main inferential tests at the group level are *t*-tests and correlations. Even though fMRI analysis always involves comparing means of different trial types by subtracting one from the other (i.e., in a pairwise *t*-test), this subtraction is always done at the subject level; at the group-level, there is usually only one data point per participant (per voxel). Thus, one-sample *t*-tests are used to compare the group mean contrast value (e.g., condition A beta–condition B beta) to the null hypothesis value of 0. Independent-samples *t*-tests are used when comparing different groups such as smokers to nonsmokers (Galvan, Poldrack, Baker, McGlennen, & London, 2011) or children to adolescents (Pfeifer, Masten, Borofsky, Dapretto, Fuligni, & Lieberman, 2009). Correlation/regression is used when there is a subject-level individual difference or behavioral measure that is hypothesized to relate to brain activity. In these types of second-level models, the correlation across subjects is estimated between the contrast value and the measure at each voxel, producing a map of where in the brain the difference between conditions is related to the behavioral measure. Although often called a brain-behavior correlation, this approach is more accurately described as a moderation or interaction because it tests for regions where the difference between conditions varies as a function of the behavioral measure. For example, some of the authors recently found that a region of the frontoparietal attention network was active overall during mindfulness meditation versus control (i.e., a main effect), and that trait mindfulness moderated activity in a separate region such that this region was more active during mindfulness meditation versus control only for highly mindful people (i.e., a task-by-trait interaction ; Dickenson, Berkman, Arch, & Lieberman, 2013).

Statistical thresholding. Each of the statistics described earlier is computed at every voxel in the brain (or within the hypothesized ROI), and some of these voxels are likely to evince a large test value by chance. For example, a typical map of contrast values contains $64 \times 64 \times 34 = 139,264$ voxels that are each $3 \times 3 \times 3 \text{mm}$, even though many of them are outside the brain or contain white matter or cerebrospinal fluid, which are not expected to show blood flow differences between conditions. (Only about 53,500 are likely inside gray matter.) Even if there is no difference between the conditions in the contrast, without eliminating these extraneous voxels and using an α -level of .05, this analysis is expected by chance to identify 6,963 voxels as “significant” (i.e., commit Type 1 error in 6,963 voxels). As the main problem is the large number of statistical comparisons, one approach would be to use a traditional adjustment such as a Bonferroni correction, but these are overly conservative (e.g., Bonferroni would suggest a voxel-wise alpha threshold of 3.59×10^{-7} in this case, which would

eliminate even the most reliable activations known to neuroimaging such as visual cortex activity during visual stimulation). Another approach is to use ROI analyses to reduce the total number of comparisons, but this denies researchers the opportunities to make discoveries that were not hypothesized. The challenge in statistical thresholding of wholebrain fMRI data is to find a way to account for multiple comparisons in a way that balances Type I and Type II error rate (Lieberman & Cunningham, 2009).

There are now at least a dozen types of statistical thresholding ranging from more conservative to more liberal, and more surely will be developed in the coming years. For example, familywise error rate (FWE) restricts the probability of finding *any* false positives, is relatively conservative, and includes Bonferroni correction and random field theory (Nichols & Hayasaka, 2003), whereas false discovery rate (FDR) controls the fraction of detected voxels that are false positives (Benjamini & Hochberg, 1995; Genovese, Lazar, & Nichols, 2002), is less conservative, and includes some simulation methods (Forman, Cohen, Fitzgerald, Eddy, Mintun, & Noll, 1995) and cluster-level correction approaches (Chumbley & Frison, 2009). There is even now a “threshold-free” method of selecting significant clusters that appears to be highly sensitive without inflating the Type I error rate (Smith & Nichols, 2009), but it is not yet implemented in most software packages. All of these methods are actually more conservative than are the typical approaches in social psychology publications (Lieberman & Cunningham, 2009). There is consensus in the field that some correction for multiple comparisons is needed – particularly for wholebrain approaches – but there is not yet an optimal solution. (Using ROIs reduces this problem, but there is simply insufficient data on many topics at this early point in the field to make strong a priori predictions about where activation is likely to be for a given psychological process.) For now, researchers and reviewers must use their judgment whether the level of thresholding is too liberal or too conservative based on the potential value of false positives relative to missed true effects.

Reporting Standards

Poldrack and colleagues have written the definitive standard for reporting a cognitive neuroscience fMRI study (Poldrack, Fletcher, Henson, Worsley, Brett, & Nichols, 2008). As this applies just as well to social neuroscience, we will not reiterate it here but instead highlight a few of the authors’ recommendations and also mention some others that are particularly relevant to social and affective neuroscience. The two main types of standards that are relevant across all the

cognitive neurosciences are those that ensure unbiased reporting of results and those that facilitate meta-analysis (see Johnson & Eagly, Chapter 26 in this volume).

No paper can report the entirety of the large corpus of data generated in an fMRI study. How does one select which analyses to report, and at which threshold? The best solution is to be as forthcoming as possible in describing all conditions in the design and all comparisons among them that are relevant to the hypotheses. If a comparison is relevant to the hypothesis, then it should be reported regardless of the results. For example, if a study on emotional reactivity included conditions with positive and negative emotional stimuli, as well as neutral stimuli, then the authors should report a set of orthogonal contrasts that completely describes the variability between the conditions. Furthermore, it is not appropriate to omit conditions that were in the design from the report because all epochs from a task must be included in the statistical model. In the emotion reactivity example, for instance, it would be inappropriate to omit the positive emotion condition from the report even if (maybe especially if) there were no brain regions that showed a significant difference between it and either of the other two conditions. Not all comparisons must be featured in a figure, but all suprathreshold clusters from all hypothesis-relevant comparisons must be reported in a table. Statistical thresholds for wholebrain analyses should be decided in advance of the study (and hence described in the Methods section) and applied equally to all comparisons. If the authors have specific regional hypotheses, then ROI analyses should be used instead with standard thresholds (e.g., $\alpha < .05$ for one ROI, with appropriate corrections for multiple comparisons for >1 ROI) and supplemented with wholebrain analyses using a different threshold. We note that the lower threshold for ROIs is only appropriate if (1) the exact size, shape, and location of the ROI is specified in advance and (2) the “selective averaging” ROI approach is used in which all voxels within the ROI are included equally in the analyses (i.e., no within-ROI search for activation).

Meta-analysis is particularly important in social neuroscience given the relatively young age of the field and the unknown Type I error rate (see Johnson & Eagly, Chapter 26 in this volume). At this stage in the science, researchers have been understandably more interested in making new discoveries than in protecting against potentially spurious ones, with the explicit understanding that true effects will emerge with replication and in meta-analysis. It is thus especially important to ensure that all activations are reported in a way that makes them accessible to meta-analysts. In addition to clearly describing each condition and reporting all contrasts among them as described previously, this

also requires reporting the coordinate system and normalization template, locating each cluster with coordinates in that system, providing the size of each cluster (in mm³ or in voxels and per-voxel size in mm), labeling each cluster with an anatomical region, and describing the anatomical localization procedure (e.g., which atlas was used). Many researchers also report putative Brodmann's areas (BAs), with the caveat that this atlas is based on cytoarchitectural structure and cannot be directly inferred from fMRI data. (For example, there are boundaries between BAs within macroanatomical gyri that cannot be identified using even the highest-resolution anatomical scans available.) A good compromise is to report Brodmann's areas and complement them with descriptive labels generated by “tedious neuroanatomy” (Devlin & Poldrack, 2007), the invaluable process of sitting down and manually identifying each locus of activation.

Finally, there are a few reporting issues that are specific to social and affective neuroscience. Because scan time is so expensive, researchers often include several different tasks within a scanning protocol. The other tasks in a protocol might be irrelevant in some fields (e.g., vision), but can be critical in social neuroscience because our psychological processes of interest can be highly sensitive to context and recent events (e.g., ego depletion effects; Baumeister & Heatherton, 1996, or priming; Tulving, Schacter, & Stark, 1982). For this reason, social neuroscientists should consider reporting the other tasks completed in the same scanning session that might affect performance on the task of interest. A related issue is participant expectations, which can also be influenced by contextual effects. For example, attributions about the cause of social rejection can alter its neural response (Masten, Telzer, & Eisenberger, 2011), and attributions can be influenced by subtle contextual information such as the likelihood of future interactions or whether the participant believes the interaction partner saw him/her. Hence, as in other types of social psychology research, these small but potentially powerful details must be included in the research report.

Electroencephalogram (EEG) / Event-Related Potentials (ERP) Methods

As noted earlier, the primary weakness of fMRI methods is their inability to track the time course of neural activity, and many theories in social and personality psychology are explicit not about where psychological processes

occur, but rather about when they occur and in which order. For example, several important models of prejudice regulation propose that automatic stereotypes and affect are activated automatically, and that controlled processes may inhibit or control this impulse (Devine, 1989). Although the processes of automatic activation and later control may be localized in different brain regions, models such as these are inherently more about time than they are about space. In contrast to fMRI, electroencephalogram (EEG) methods allow for millisecond-by-millisecond recording of the electrical activity caused by neural activity. The temporal accuracy of EEG is possible because changes in postsynaptic potentials (when neurotransmitters bind to receptors) following neural activity create immediate changes in measureable electrical signals on the scalp (Buzsaki, Traub, & Pedley, 2003). The electrical signals, unlike the BOLD signal, are a direct measure of neural activity *at the time that activity occurred*. Thus, EEG recordings contain the exact information that is lost in fMRI recordings – when the processes occur, and in what order. As models of social cognition begin to more fully embrace dynamical systems models of cognition rather than rigid dual-process models (Cunningham & Zelazny, 2007; Richardson, Dale, & Marsh, Chapter 11 in this volume), such information will be essential for providing more nuanced information regarding the interactions among multiple component processes.

Yet, this temporal precision gained with EEG comes at a cost. Whereas fMRI can tell us where a process is likely to have occurred, but not when, EEG methods can tell us when a process occurred but not where. Multiple factors contribute to this loss of information. First, neural activity is summed across multiple foci (or generators) inside the brain by the time this information reaches the scalp. The signal coming from a right anterior electrode is not necessarily, or even primarily, coming from the right prefrontal cortex. A left parietal area may be primarily responsible for this effect, if the electrical activity is oriented toward the right frontal regions. Furthermore, if multiple regions are involved in a process, these processes are summed across all the electrodes. Second, the scalp itself blurs the signal in space such that the electrodes, even when using a high-density system (128 or more channels), do not provide independent signals. Although new methods for determining the spatial location of EEG signals are being developed (Pascual-Marqui, Michel, & Lehmann, 1994; Scherg & Ebersole, 1993), they suffer from the fact that a near-infinite number of solutions can be generated when attempting to isolate the location of the EEG signal – there is no unique solution (Plonsey, 1963). Thus, EEG provides quantitatively different information about neural activity than fMRI does. Because we must

choose to sacrifice knowing where for when, or when for where (leading to something analogous to the Heisenberg Uncertainty Principle in physics), the choice to use fMRI or EEG should come from theoretical questions rather than from ease or cost.¹

Study Design

Whereas the EEG is the raw time-series of electrical activity recorded (see [Figure 7.8A](#)), an ERP is the averaged neural activity associated with a particular trial type. Specifically, an ERP reflects the averaged time-locked signal following neural activity. Given the oscillatory nature of EEG recording, ERPs typically have multiple deflections, some with a more positive voltage than baseline and some with a more negative voltage than baseline. Unlike fMRI, where a positive signal reflects activation and a negative signal reflects deactivation, the direction of the signal (positive or negative) is relatively meaningless for EEG. Rather, the strength of the absolute effect, positive or negative, is meaningful. To calculate an ERP, one creates *epochs* (time windows that correspond to specific trials) from the raw time-series data to particular time points (e.g., when the stimulus was presented, or when a button press was recorded) and lines up all the EEG data in time for each of the predefined trial types. These time points are then averaged to create a *grand average* for each trial type for each participant that can be compared across condition (See [Figure 7.8B](#) for an example). Although this averaging can distort the individual trial-to-trial or the participant-to-participant variability, this is important because it increases the signal-to-noise ratio and allows for meaningful comparisons to be made. Deflections – or ERP components, or waveforms – are typically labeled with a letter and a number. The letter corresponds to whether the deflection was positive or negative, and the number corresponds to when the component occurs. The numbers can reflect either the number of deflections since baseline (e.g., a P1 would reflect the first positive deflection and an N3 would reflect the third negative deflection) or the time in milliseconds from the baseline (e.g., a P100 would be a positive deflection around 100ms after baseline, and an N170 would be a negative deflection 170ms following baseline). Other ERPs have specific names, such as the ERN (error related negativity) that follows behavioral mistakes (Gehring & Fencsik, [2001](#)), or the LPC (late positive component) associated with explicit recognition memory (Rugg, Schloerscheidt, Doyle, Cox, & Patching, [1996](#)).

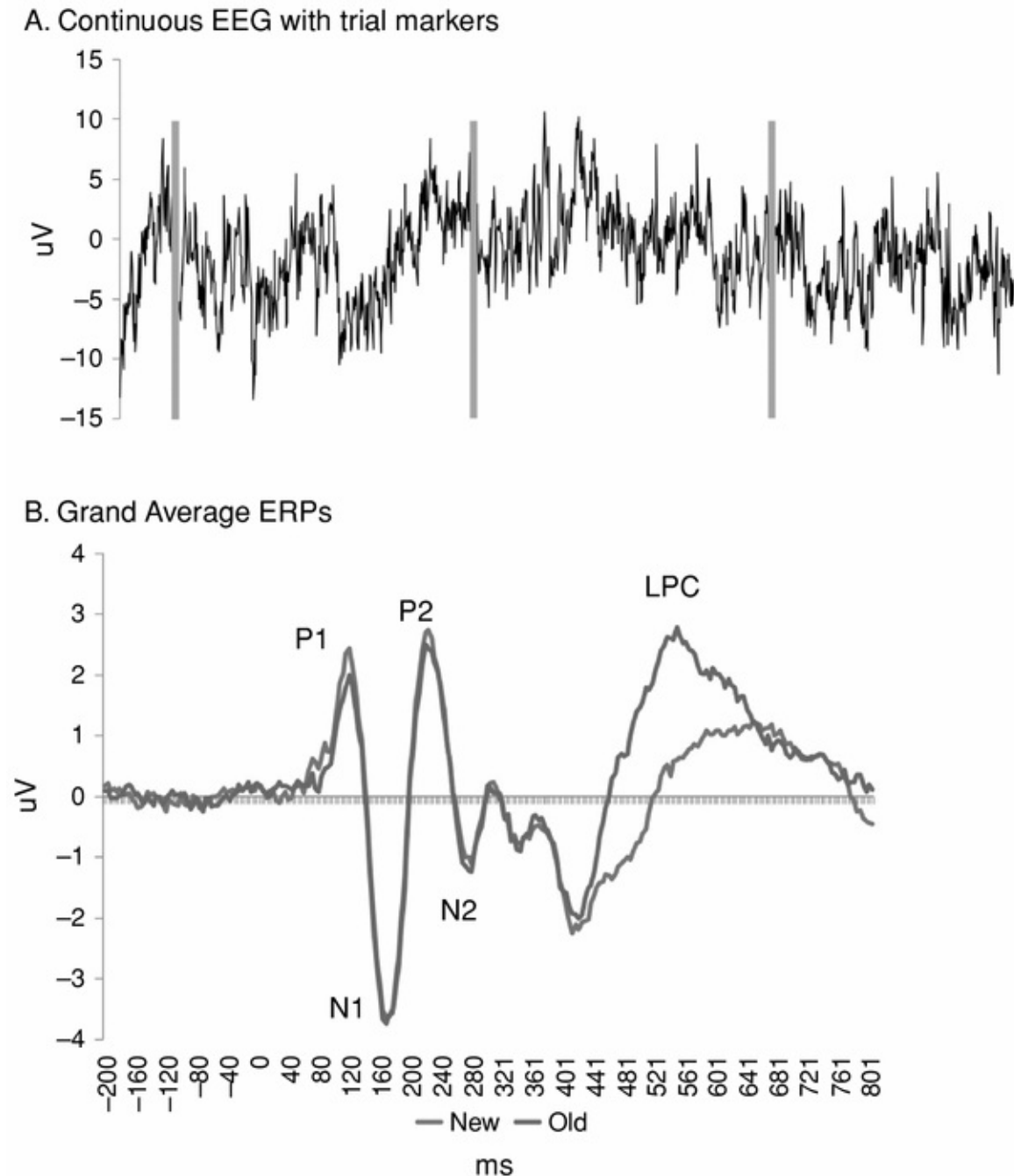


Figure 7.8. (A) Continuous EEG data sampled at 250Hz. (B) ERP components for new and old memory judgments. The waveforms shown are averages across many trials of each type, time-locked to the onset of the trials (depicted as vertical bars in A). ERP components are labeled for positive vs. negative deflections. Unpublished data for figure courtesy of Per Sederberg's Computational Memory Lab at <http://memory.osu.edu>.

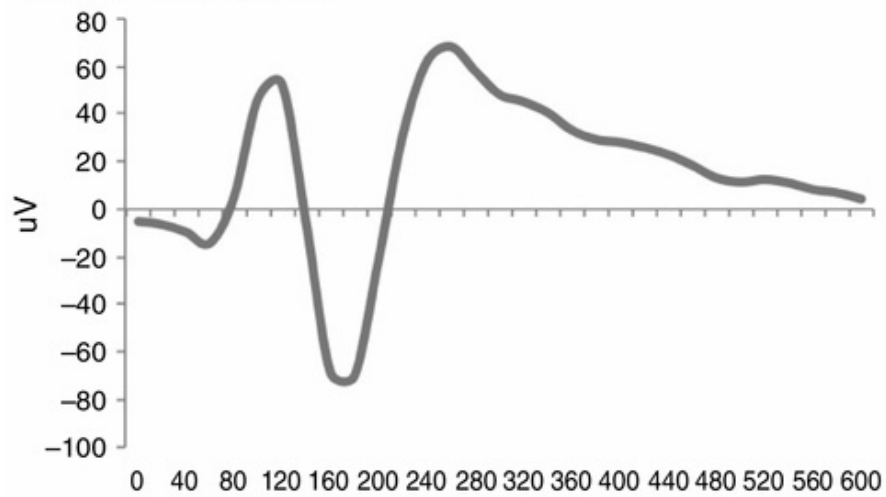
In one of the first social psychological studies using EEG, Cacioppo, Crites, and Gardner (1996) identified an ERP associated with valenced stimuli presented in an emotionally incongruous context. A series of valenced stimuli were presented before a critical stimulus that was of the same or different valence, and EEG signals that differentiated the stimuli presented in congruous versus incongruous contexts were examined. Cacioppo and colleagues identified a particular type of wave form termed a *late positive potential* (LPP), occurring when participants saw a stimulus that was incongruous with a context; in these studies, a negative stimulus in the context of positive stimuli, or a positive stimulus in the context of negative stimuli. The amplitude of the LPP wave in these studies was shown to vary as a function of the degree of difference between the valence of the stimulus and the valence of the context in which it occurs. For example, when presented in the context of positive stimuli, a strongly negative stimulus will result in a larger LPP than a mildly negative stimulus (Cacioppo et al., 1996; Cacioppo, Crites, Gardner, & Berntson, 1994). The LPP associated with evaluative incongruity is widely distributed across scalp electrodes but is more pronounced over posterior (parietal) scalp regions than over frontal sites. There is also evidence that the amplitude of this posterior LPP is greater over the right hemisphere than over the left for both positive and negative stimuli presented in an incongruous evaluative context (Cacioppo et al., 1996). Researchers using this paradigm have shown that the posterior LPP is evident when participants are making both evaluative and nonevaluative judgments, suggesting that evaluative incongruity may be detected automatically (Cacioppo et al., 1996; Ito & Cacioppo, 2000; see also Crites & Cacioppo, 1996). LPPs show a negativity bias in that they are typically larger for negative stimuli in a positive context than positive stimuli in a negative context (Ito, Larsen, Smith, & Cacioppo, 1998), and the degree of hemispheric asymmetry (right greater than left) is greater for negative stimuli as well (Cacioppo et al., 1996).

Like fMRI studies, EEG/ERP studies need to examine changes in neural activity as a function of a presumed change in perceptual or cognitive processing. The brain is always active, in the sense that neurons throughout our brains fire spontaneously even when we are resting, and so it is only possible to quantify changes in brain activity. To examine these changes, it is necessary to have at least one within-subject condition that varies as a function of some psychological state. This variation can come from the manipulation of a categorical variable (pleasant vs. unpleasant photographs) or by manipulating a variable continuously (degrees of valenced photographs). As with all

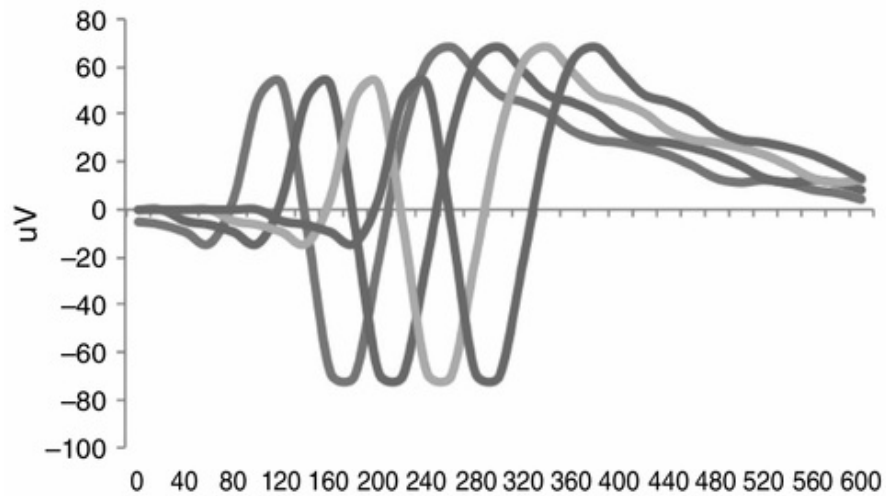
neuroscience methods and as discussed previously, it is important to ensure that either only one variable is being manipulated or that appropriate control conditions have been added to statistically control for potential confounds. Further, because of the relatively low signal-to-noise ratio for neuroscience methods, it is necessary to have many trials of data. Because of these similarities, rather than repeat those design issues here, we highlight in this section the unique challenges that emerge when collecting ERP data and provide suggestions about how to optimally design studies to minimize these concerns (for a more detailed and exhaustive discussion of EEG methods, see Luck, 2005).

The goal of ERP data analysis is to characterize a neural response in time. As noted earlier, the advantage of the EEG method is its remarkable precision in detecting when a particular neural response occurs. This precision, however, means that additional assumptions must be made during the study design and analysis, and it is possible that some questions simply cannot be answered using ERP methodology. The fundamental issue is that to average multiple trials into a single ERP waveform, we must assume that the processes occur at the same time point. For example, an early visual component in the occipital electrodes will be found if, for most trials and for most participants, there is a positive deflection approximately 100ms following stimulus presentation (see Figure 7.9A). However, note what happens if for some trials the effect is offset by intervals of 40ms (Figure 7.9B). When these trials are averaged, the positive and negative deflections of the underlying neural process cancel each other out, leaving no effect whatsoever (see Figure 7.9C).

A. Process locked ERP



B. Onset shifted ERP



C. Grand average from onset shifted ERP

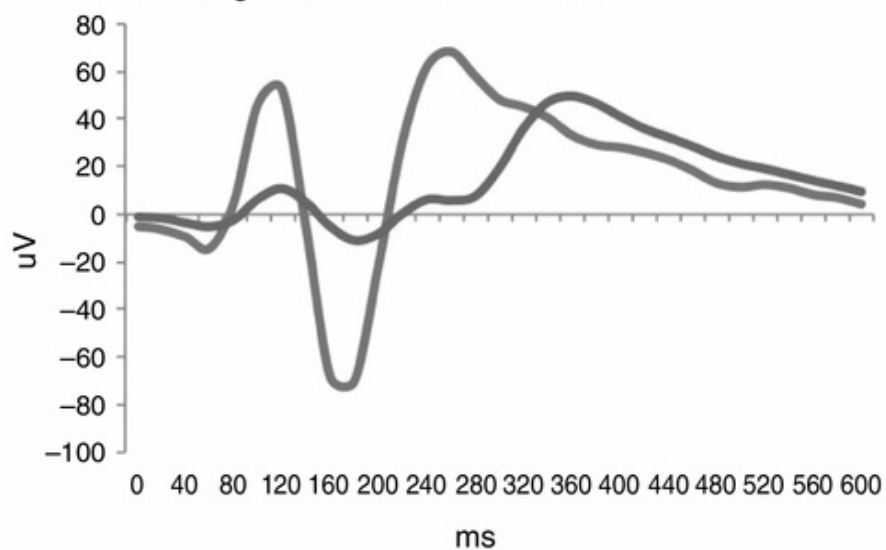


Figure 7.9. (A) Simulated ERP data with a P100 and N170 component associated with face processing. (B) Simulated data moving the onset. (C) Grand average data combining the ERPs with different onsets compared against the modal ERP.

What these examples illustrate is that not only must it be assumed that the brain regions that produce the effect are the same for all trials and all participants, but that the timing of the brain activity must also be consistent. Because of this, it should not be surprising that many, if not most, ERP studies have examined low-level perceptual process or motor responses. In both cases, it is more likely that the neural responses will “line up” in exactly the same way. For a perceptual process, there is a ballistic sequence of events that needs to automatically decode the stimulus properties to create a representation of an object. The ERPs following the stimulus presentation are likely very consistent from trial to trial for the first hundreds of milliseconds. For a motor response, we know when the response occurred, so we can look at the ERPs that occur just before or just after the response. Again, we can assume that these responses are likely consistent for hundreds of milliseconds before or after the response.

Although the time locking to either a stimulus or response allows to understand the components of many automatic processes, this presents challenges for the study of some constructs within social and personality psychology. For processes such as Theory of Mind, social rejection, or intertemporal discounting, it is unlikely that there are automatic sequences of processes events that (1) unfold in exactly the same way for all trials within subject (not even considering between subject difficulties), and (2) that we can a priori identify the temporal onset and offset of processing. For example, when deciding whether someone knows where an object is hidden, there is no reason that a person cannot consider another person's intention at 500ms, or 1,200ms, or even 3,000ms after the information is presented. Further, for a complicated decision, people can consider the various components of the information in different orders – in a gambling task, a person in one trial can consider the amount of money that can be earned (\$50 vs. \$1) before the probability of winning (60% vs. 20%), and in another trial can consider the probability before the magnitude. In this example, if amount of time spent considering the magnitude may vary as a function of the size of the potential win (i.e., if a participant spent more time thinking about \$50 than \$1), then these timings would further become blurred in time. Thus, as cognitive flexibility and

reflective processing increases, we should see an exponential increase in the blurring of ERPs. Considering the importance of time locking (to the millisecond!) neural processes, it should not be surprising that most ERP effects discussed in the literature occur within the first hundreds of milliseconds before or after the time locked trial.

Data Collection

Among the most important aspects of EEG/ERP data collection is to collect as clean data as possible. EEG signals from neural activity are small (less than 1/100,000th of a volt), and have a very low signal-to-noise ratio, so maximizing the signal and minimizing the noise components are important. Before all else, it is important to prepare the room for data collection and remove as many of the extraneous sources of electrical activity as possible. Many people build rooms specifically for EEG recording, which isolate electrical signals – removing computers from the rooms, using DC instead of AC lighting, and so on. The fewer sources of electrical activity in the testing room, the better the EEG recordings of neural activity will be. Next, it is important to get the best signal from the EEG cap. To do this, either a gel or a saline solution is placed under the electrode to decrease electrical resistance (impedance) between the electrode and the scalp. Reducing the impedance of the electrical recording allows for better recording of the neural activity and reduces the recording of environmental noise. Gel solutions allow for lower impedances (better signal), but are messy and take time to apply. Saline-based solutions typically cannot get as low impedance as gels can, but are less messy and faster to apply. If time is not an issue, gel tends to be preferable to get a better signal, but because application can take more than 30 minutes for some high-density systems, a saline solution can be preferred if working with certain populations, such as children, where there may be limited patience or time to apply the gel.

After preparing the electrodes for recording, it is necessary to determine a reference electrode. EEG is not an absolute signal, but rather the difference in voltage between two recordings. Because the activity at the reference electrode will be subtracted from all recordings, it is desirable to select a reference that is as isolated from the signals of interest as possible. Two of the most common reference locations are behind the ears (mastoid reference; recording from both the right and left and averaging these together to create a reference) or on the nose. The selection of reference location is important because it will partially determine the shape and location of observed ERP signals. In other words, a P1

recorded in an occipital electrode using a mastoid reference is *not* the same as a P1 recorded from the same electrode using a nose reference, because they examine the voltage differential across different locations and distances. When comparing across studies, or when attempting to replicate previous work, it is critical to note the location of the reference (see Joyce & Rossion, 2005, for an example).

Because precise stimulus timing is critical for ERP data analysis, the precision of the stimulus presentation package and the calibration of the stimulus monitor and response box are far more important than in fMRI data collection. For example, software for experimental presentation varies in the quality of the precision of timing, which can interact with the particular hardware that it is being run on. Further, computer hardware matters far more for ERP studies than fMRI does. One has far less control of the presentation of visual stimuli on an LCD monitor than a CRT monitor, and most keyboards have an error of response in the 50ms range. When time locking to a neural response that occurs in the order of tens of milliseconds, these errors can add substantial noise and variability to the recording.

Data Averaging, Cleaning, and Preprocessing

To calculate the grand average ERP for each electrode, a marker file must be obtained that precisely (to the millisecond) codes when events occurred in the continuous EEG. These markers can represent when a stimulus was presented (and/or when it was removed from the screen), when a response was made, or any other event that has psychological meaning. It is always better to include too many than too few markers. When computing the grand average ERPs, it is necessary to determine which of these markers will be used to align the individual trial data in time. If interested in perceptual processing, the stimulus marker will most often be used. If interested in motor responses or decision processes, the response marker will most often be used. It is important to note that the resulting grand average ERPs time locking to the stimulus or the response will be completely different, as the stimulus marker does not perfectly predict the timing of the decision processes and motor responses, and the response marker does not perfectly predict the timing of the stimulus presentation. Indeed, because of the blurring of ERPs discussed in the previous section, very little motor-or decision-related activity will be found when time locking to the stimulus, and very little perceptual processing will be found when time locking to the response. Because the temporal dynamics are critical in ERP

studies, decisions about time locking should be made relative to how theory predicts consistency in temporal responses across trials and participants.

Before averaging, it is important to clean the data to ensure that most of the variability is attributable to psychological processing rather than extraneous sources. This is done through two primary means: data filtering and trial rejection. EEG data is typically bandpass filtered prior to calculating the individual trial epochs and the grand averages. Typically, a 0.1 to 0.30Hz bandpass filter (only frequencies between 0.1 and 0.30Hz are retained) is applied to the data to remove most extraneous sources of noise. Tighter filters can be used for particularly messy data, such as when uncooperative participants or children are run. Ideally, however, you would want to keep as much data as possible, or at least be confident that nothing you exclude could be meaningful, task-related neural activity. Following bandpass filtering, trials that have too much artifact are typically removed from the analysis. The most common artifacts are eyeblinks, eye movements, and head motions. These artifacts create huge changes in voltage that dwarf the voltage changes that come from neural activity. Luckily, these artifacts are well characterized and can be deleted manually by examining each individual epoch, or can be detected using automated algorithms. More recent developments have allowed for the estimation of these artifacts (e.g., epoching ERPs to participant blinks) that may allow for artifact correction rather than removal. Eyeblinks and eye motion can be reduced by allowing participants' periods of time when they can blink frequently throughout the task and by designing tasks that require limited visual search.

Data Analysis.

Assuming a sampling rate of 500Hz, and a 128-channel EEG system, 64,000 data points are collected per second. Thus, if an epoch is defined as the 1,000ms following a marker, we are left with tens of thousands of data points per condition – an amount of data that leads to a serious problem concerning the appropriate ways to correct for multiple comparisons. Assuming only two conditions, if a researcher tested each time point at each electrode for these two conditions, 3,200 false positives would be found using a $p < .05$ alpha level.

Luckily, EEG data is not independent in either the temporal or spatial domains, and several conventions allow for a relatively straightforward way to reduce the number of multiple comparisons. Specifically, researchers often examine only one or two meaningful ERP components and average the data

across multiple time points to get a single estimate of the ERP effect. For example, if one is interested in the P1 visual component, because this effect typically occurs around 100ms following stimulus presentation, it can be quantified as the mean signal averaging from 90 to 110ms. Similarly, the N170 visual component can be isolated as the mean signal from 160 to 180ms. By focusing on only these two components (in this case known to be involved in face processing), a researcher can already reduce the data from 64,000 time points to 256. Not only does this reduce the number of comparisons, but because the EEG signal itself is not independent, it also helps increase the signal-to-noise ratio by averaging data that presumably capture the same information.

In addition to collapsing in the time domain, it may be beneficial to collapse in the spatial domain. In the faces-processing example earlier in the chapter, only some electrodes may be meaningful to examine. For example, given previous research, posterior electrodes are more likely to provide more meaningful signal than more anterior electrodes when examining visual processing. Because of this, a researcher can focus on a subset of electrodes (perhaps 20 in the left hemisphere and 20 in the right hemisphere), thereby reducing even further the number of comparisons. Given a strong enough a priori hypothesis, there is no reason to go beyond the examination of a single electrode. For example, Amodio, Harmon-Jones, Devine, Curtin, Hartley, and Covert (2004) focused on how people respond when making errors associating prejudicial objects such as guns with African-American faces. Specifically, participants were more likely to classify a tool as a gun following an African-American face than a Caucasian face. Because they were interested in error processing, Amodio and colleagues focused on one particular ERP, the event-related negativity (ERN; Gehring & Fencsik, 2001; van Veen & Carter, 2002), which was associated with cortical activity after detecting errors. Because the ERN was well characterized in previous years, they were able to focus on a single frontocentral electrode (the Fcz) for all analyses and avoid problems of multiple comparisons. Similarly, researchers interested in face processing will tend to focus on the P100 and the N170 components time locked to the stimulus presentation, as these components have been shown repeatedly in the literature to be associated with face processing, and they may focus more on posterior lateralized electrodes (Bentin, Allison, Puce, Perez, & McCarthy, 1996). Thus, as in fMRI, the ways in which one reduces the data are dependent on the specific question being asked and the previous literature in this domain.

Once the data reduction strategy has been determined, ERPs are extracted from the single-subject grand averages. There are two primary methods to

quantify the size of the ERP component. One can determine the peak amplitude by identifying the highest (or lowest) point and labeling that the size of the ERP effect. Although this approach has intuitive appeal, it can underestimate the size of an ERP for subjects (or conditions) where there is greater variability in temporal dynamics of the ERP component. For example, if in one condition the peak occurs at 100ms, ± 10 ms, and in another condition the peak was at 100ms, ± 20 ms, the fact that the peaks are more spread out in the second condition would reduce the observed average effect size even if the trial-by-trial peak amplitudes were identical. To circumvent this problem, it is preferred to calculate the mean amplitude across the entire ERP time window (the mean of time points 90–110ms). This method captures the average amplitude and therefore the contributions of each of the independent trial amplitudes to the grand average.

If multiple electrodes are investigated simultaneously, it is standard to include electrode site and each of the ERP components as factors in an ANOVA design. So, if someone were interested in the P100 and the N170 mentioned earlier to be sensitive to faces, and wanted to compare upright versus inverted faces, they may have to compare 20 left hemisphere and 20 right hemisphere posterior electrodes on a high-density EEG system. To analyze, this would result in a 2 (upright/inverted) X 2 (P100/N170) x 40 (electrode) MANOVA. Because the electrode sites are not independent from one another (here, the right electrodes are likely to be artificially more correlated with one another than they are with the left electrodes), it is necessary to perform a Greenhouse-Geisser correction any time that more than two electrodes are entered into the model. Because of this, if one believes that the right and left effects are for the most part isomorphic, one can simply average these effects to create a 2 (upright/inverted) X 2 (P100/N170) x 2 (left/right) MANOVA.

Analysis of Continuous EEG

Up to now, we have focused on the examination of single ERPs following stimulus presentation or a response. Although these ERPs are useful for investigating automatic processes, they have limited utility for examining cognitive processes that may not follow a stereotyped temporal pattern. Although this is a limitation of ERP data collection process and analysis, the brain dynamics associated with more complex cognitive processes are recorded in the continuous EEG signal and can be recovered using different methods. One such technique is to compute the power in different frequency bands across several seconds (and often minutes) of EEG data. Power simply reflects the

degree to which there is a large amount of activity within a particular frequency band. Greater power in these different bands at different electrode sites reflects different degrees of processing, *in general*, across tasks or people.

One of the best examples of using continuous EEG to examine affective processing comes from the work on frontal asymmetries for approach or avoidance motivation. Following the literature on emotional processing following brain injury, Davidson and his colleagues predicted that the right hemisphere may be more associated with negative affect and the left hemisphere with positive affect. To test this idea, Tomarken, Davidson, Wheeler, and Doss (1992) recorded EEG while participants were at rest, with the hypothesis being that greater activation in the right compared to left hemisphere would be found *across the entire resting epoch* for people who were predisposed to depression. To analyze the continuous EEG across the entire time window, the data were decomposed into different frequency ranges and the power of these ranges was estimated. By comparing the power estimates from right and left hemisphere, they found that differences in the alpha band (8–12 Hz) predicted affective style. Specifically, greater right-sided power is associated with greater negative symptoms. More recent research has suggested that these frontal asymmetries may be more associated with approach (left) versus avoidance (right) motivation (Harmon-Jones, Lueck, Fearn, & Harmon-Jones, 2006). Regardless of the specific mechanism, frontal EEG power asymmetries at rest have been shown to predict depression, emotion regulation ability, and general well-being (Davidson, 1988; Jackson, Mueller, Dolski, Dalton, Nitschke, Urry et al., 2003; Urry, Nitschke, Dolski, Jackson, Dalton, Mueller, Rosenkranz, Ryff, Singer, & Davidson, 2004).

Eight Conceptual Issues in Social Neuroscience (and How to Think Clearly about Them)

1. Forward and Reverse Inference

In neuroimaging, typically psychological processes are the independent variables and neural activity is the dependent variable. The *forward* direction of logical inference, therefore, is from the psychological to the neural: if psychological Process A is engaged and Region X is activated, then we can conclude that Region X is involved in Process A in some way (though not necessarily causally). Neuroimaging works this way because the methods were optimized to

answer questions about localization of function in the cognitive neurosciences, for example, to identify regions in the brain that are responsive to motion perception. The vast majority of studies in social neuroscience still take the approach of manipulating a psychological process to the end of localizing the neural systems that are recruited during that process, and the methods described thus far in the chapter are valid inferential tools for this purpose.

However, it is tempting to do the reverse by inferring the presence of a psychological process based on neural evidence (e.g., concluding that motion perception must have been engaged based on activation in a particular area). This is known as *reverse inference*, and is often – albeit not always – an instance of the logical fallacy of affirming the consequent (Poldrack, 2006). The fallacy is that “ $B \rightarrow A$ ” does not logically follow from “ $A \rightarrow B$ ” when the relationship from psychological process A to neural activation B is not unique (i.e., when process A is not the only one that is associated with activation in B). And nearly all of the time, it is not. Indeed, the mapping from mental processes to neural activations seems to be *many-to-many* in that most brain regions are involved in many psychological processes, and vice versa. Thus, the only logical conclusion that can be made based on observing activation B is that one of the many psychological processes that have been observed to be associated with B (or one that is not yet known to be) was engaged. We assume you agree that this is not a satisfying answer.

What is needed is a comprehensive database of social neuroscience data that tracks the mapping between cognitive processes and brain activation in order to estimate the probability of a psychological process given activation in a region (Poldrack, 2010). Bayes's Theorem provides a way to do this, but requires some base rate frequency information. Bayes's Theorem states that to make the reverse inference that a mental process was engaged based on a neural activation, it is necessary to know the probability of activation in that region during the mental process, the probability of the mental process being elicited in the given task (which should be close to 1 if the task is valid), and the probability of activation of that region across all mental processes. In cases where the mental process is highly specific to a given region – that is, the probability of activation is high in the process of interest but low in others – reverse inference is more likely to be valid. However, in cases where the mental process is not specific to a given region – that is, the probability of the activation is high regardless of the mental process – reverse inference is less likely to be valid. The insular cortex is an example of a brain region falling into this latter category, as activation there is observed across a broad variety of tasks (Kurth, Zilles, Fox, Laird, & Eickhoff,

2010). Thus, it is not valid to infer a specific mental process based on observed activation in the insula alone – a mistake that has been made at least once in the New York Times (Poldrack, 2011).

Fortunately, researchers increasingly recognize the need for tools to enable them to gauge how specifically a region is involved in a task. There are several such databases including NeuroSynth.org (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011), the Cognitive Atlas (Poldrack, Kittur, Kalar, Miller, Seppa, Gil et al., 2011), and the NeuroLex neuroscience lexicon (Bug, Ascoli, Grethe, Gupta, Fennema-Notestine, Laird et al., 2008). The use of these systems is not yet standard in the field and many of them are still in beta stage, but we anticipate that they will be commonly used within the next few years to support reverse inference claims with Bayesian data.

A second way to deal with reverse inference is to focus on networks of regions rather than isolated regions. Although a given region may show activity in response to multiple psychological demands, a set of brain regions may co-activate under a more selective set of psychological conditions. For instance, ventromedial PFC is a commonly activated region in a number of qualitatively different tasks. When seen alone, it can be difficult to interpret. However, when seen in conjunction with ventral striatum and ventral tegmental area, it is more likely that this three-region network indicates some form of reward processing or valuation compared to when any one of those regions is seen alone.

2. Spuriously High Correlations?

Social neuroscience has recently received criticism for reporting strong correlations between brain activation and individual differences (e.g., behavioral or self-report measures) that seemed to some to be “likely entirely spurious” (Vul, Harris, Winkielman, & Pashler, 2009). The correlations in question are indeed high, but the critics are incorrect that this implies the correlations must be meaningless. The flaw in their argument was misidentifying the source of the high correlations to be circular analysis (Lieberman, Berkman, & Wager, 2009); the high correlations are actually a result of the stringent thresholding procedures used to protect against Type I error. (We note that this criticism applies to cognitive neuroscience in general, which uses identical methods as those described here, but the original critique focused on social neuroscience.)

The main assertion of the critics is that many correlations in social neuroscience are the product of circular statistics, or “double-dipping,” and are therefore not valid (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). The

claim is that some scientists were analyzing their data by first searching the whole brain for voxels that correlate with a measure, and then were selecting only those voxels that correlated highly with the measure for entry into a *second* analysis. We agree with the critics that such a second round of analysis would be circular and would not produce meaningful results. However, the problem with their critique is that nobody (or perhaps very few) researchers actually analyze their data in this way. Correlations in social neuroscience are generally conducted as described in the earlier section on second-level models: a wholebrain search is conducted (i.e., at every voxel) for correlations between contrast estimates and an individual differences measure, and those voxels surviving the multiple-comparisons threshold are reported. There is exactly one inferential step – the wholebrain search – and the reporting following that step is merely descriptive, regardless of whether a single voxel or an average across a cluster of voxels is reported. We note that the procedure for wholebrain correlations is exactly the same as for wholebrain condition effects (e.g., comparisons between two conditions), except the statistic computed at each and every voxel is a correlation instead of a *t*-test. However, the correlations remain strikingly high – one study reported a correlation of 0.96 between activation in the middle frontal gyrus and a reaction time measure (Sareen, Campbell, Leslie, Malisza, Stein, Paulus et al., 2007). What can explain this if not double-dipping? The cause is small sample sizes combined with multiple comparisons procedures that rely on a voxel-wise threshold. Supposing the common voxel-wise threshold of $p < .001$ and a sample size of 15, the corresponding correlation threshold (i.e., the lowest r value that could be considered significant) is 0.73. For a threshold of $p < .0001$, the threshold r is 0.82. And those correlations are the *minimum* values that can be reported with those thresholds – the tails of the distribution of values can be much higher in the presence of a true effect. The correlation values in the voxels that contain a true effect will be distributed around the true correlation according to sampling error, and one or more voxels in that population that emerge with a very high correlation (e.g., if the true correlation is 0.7 in two voxels, one might be observed at $r = 0.5$ and another at $r = 0.9$). Thus, no misapplication of statistical procedures is necessary to explain the high correlations – they are a direct result of the thresholding procedures that explicitly limit researchers to report only high correlation values.

The benefit that has resulted from this critique and the responses to it is an increased awareness of the limitations of wholebrain statistics, or at least to the limitations of the thresholding procedures necessitated by wholebrain statistics. The first lesson is to use larger samples. The relationship between a p -value

threshold (e.g., $<.001$) and a statistic threshold (usually a t or r value) changes as a function of the degrees of freedom at the group level (Figure 7.10a). For a given statistical threshold, the Type I error rate decreases dramatically as the sample size (and thus the degrees of freedom) increases, particularly until $N=20$ (Figure 7.10b). The second lesson is that our raw statistics values are not unbiased measures of effect size (nor were they intended to be) because of the thresholding procedures (Lieberman et al., 2009). Only after an effect is detected at all (which is the intention of the procedures) can an effect size be estimated for an entire functional or structural brain region by examining the correlation across that entire area. Another lesson is to use a priori ROIs to compute correlations when possible (e.g., Berkman & Lieberman, 2010), which can eliminate the need for multiple comparisons corrections in the first place and simultaneously provides effect *detection* (i.e., present or absent) and effect size *estimation*. Finally, as is always the case with empirical findings, only results that have been independently replicated should be considered reliable.

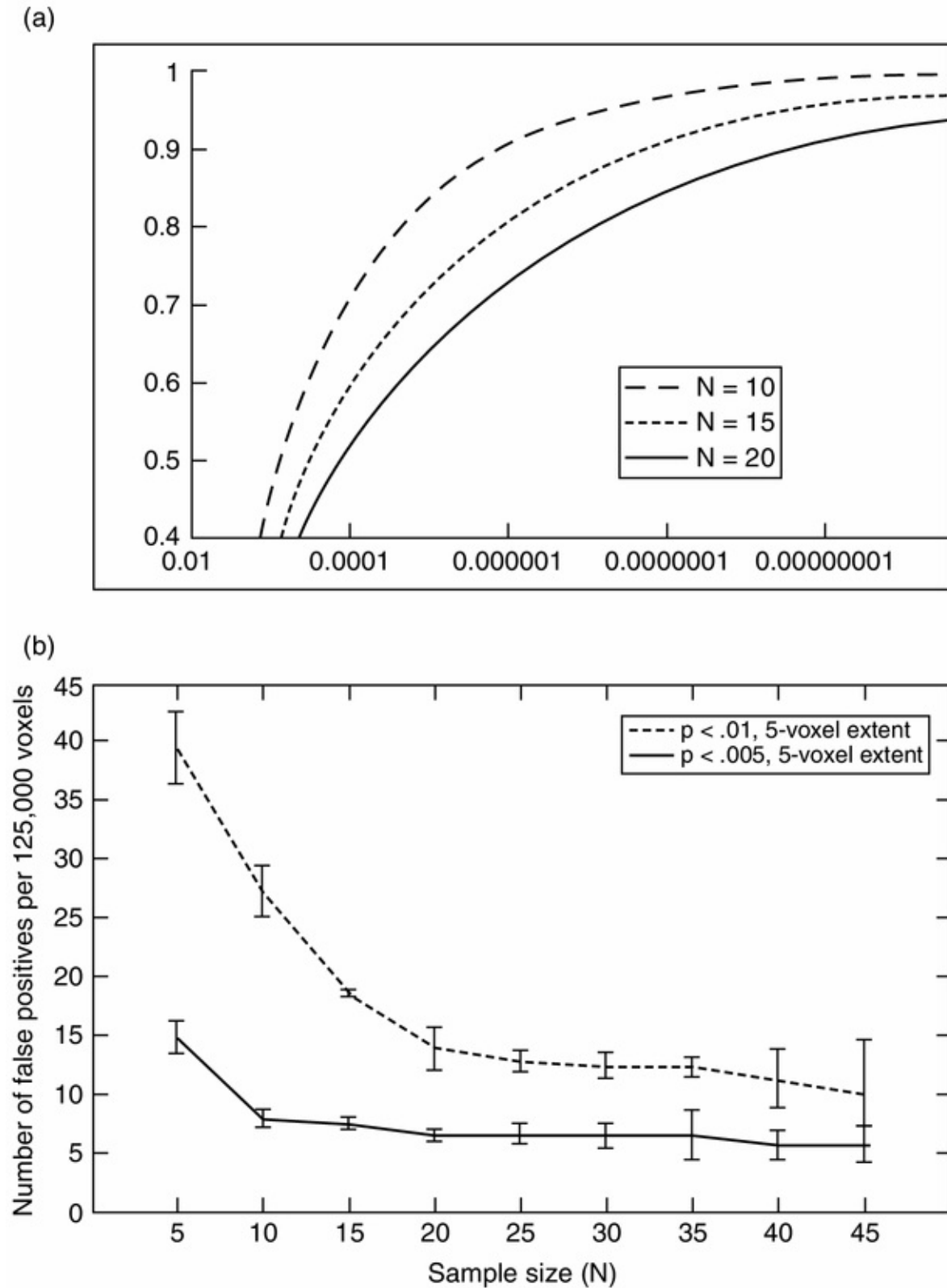


Figure 7.10. The relationship between p -value threshold and statistic threshold (e.g., correlation r) is moderated by sample size. (A) The t -distribution dictates

the relationship between an alpha level and the t (and thus r) cutoff. It becomes less conservative at larger sample sizes. (B) Simulation results plotting the number of false positive clusters as a function of sample size (N) at two different thresholds assuming no true effect. Note the steep decline in Type I error rate until N reaches ~ 20 , and the effect of the threshold in the difference between the two lines.

3. Experimental versus Ecological Validity

In any psychology experiment, there is a trade-off between experimental control of extraneous third variables or outside influences and ecological validity in realistically modeling complex human thoughts, feelings, and behaviors (Brewer & Crano, [Chapter 2](#) in this volume). This tension is present in social neuroscience as much as, or even more so than, in social psychology because the cost of null effects is quite high in terms of money and other resources. On the one hand, a key advantage of neuroimaging relative to other dependent measures is the ability to examine putative neural mechanisms of psychological processes. The best way to leverage this advantage is to examine as “pure” a process as possible by carefully controlling everything else – internal validity. On the other hand, a core aim of social neuroscience is to understand the brain systems that support real-life thoughts, feelings, and behaviors, and these cannot always be broken down into a single mental process that can be easily manipulated in isolation – ecological validity. To be realistic, we must also relax some experimental control. What should the balance be in social neuroscience?

In our view, the field of social neuroscience to date has been biased toward ecological validity, in part because one advantage of wholebrain neuroimaging is that it can assess multiple processes in disparate brain regions simultaneously, and in part as a way of differentiating itself from cognitive neuroscience. There are several good reasons that justify this bias. One reason is that social neuroscience is not concerned merely with process but also with outcome – beliefs, emotions, and behaviors that occur outside of the scanner (Berkman & Lieberman, [2011](#)), such as relationship quality or addiction relapse. Outcomes like these cannot generally be explained by one process alone, so complex scanner tasks are needed to begin to get a handle on which brain systems contribute to them and how. Another reason is simply the relative youth of the field compared to other neuroscience fields. Social neuroscience is still in the discovery phase – the very first neuroimaging experiments in a number of areas (e.g., empathy, cognitive dissonance, emotion regulation) were conducted only

within the last few years. Zeroing in on processes first requires some knowledge about which processes to examine. For example, researchers needed to conduct preliminary studies on the broad network involved in thinking about other people and their mental states (Frith & Frith, 2001; Gallagher & Frith, 2003) before breaking that network into its component pieces (e.g., Saxe & Powell, 2006; Spunt & Lieberman, 2012). We believe that, at this point, social neuroscience can make its most important contributions to psychological science by bringing real psychological phenomena into a neuroimaging environment where their corresponding brain systems can begin to be explored.

To the extent that experimental control is sacrificed for ecological validity, the degree of ecological validity should be as high as possible. It can be tempting to simplify an experimental paradigm to a pallid and repetitive response time task, and some believe that psychology in general has been guilty of yielding to this temptation too often (Baumeister, Vohs, & Funder, 2007). Social neuroscientists have embraced the importance of creating an engaging, realistic scanner experience, which partly explains the popularity of certain tasks that reliably evoke strong feelings of, for example, rejection (Williams, Cheung, & Choi, 2000), fairness/unfairness (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003), and disgust (Ochsner, Bunge, Gross, & Gabrieli, 2002). This same desire for realistic neuroimaging experiences has motivated researchers to think outside of the task itself and manipulate the imaging environment by bringing in interaction partners either physically (Coan et al., 2006) or digitally (Redcay, Dodell-Feder, Pearrow, Mavros, Kleiner, Gabrieli, & Saxe, 2010), or manipulating the temperature of tactile stimuli during the scan (Kang, Williams, Clark, Gray, & Bargh, 2011). We greatly anticipate seeing clever new ways for researchers to continue to bring the world into the scanning environment in the coming decades.

4. What Counts as Replication in fMRI and EEG?

Though test-retest reliability can be quite high both within scanner (Friedman, Stern, Brown, Mathalon, Turner, Glover et al., 2008) and between scanners (Casey, Cohen, O'Craven, Davidson, Irwin, Nelson et al., 1998), there can be considerable variability in contrast-to-noise ratio, percent signal change, and spatial normalization across scanners and individuals for a variety of reasons that are beyond the scope of this chapter but are discussed extensively elsewhere (e.g., Bennett & Miller, 2010; Jovicich, Czanner, Greve, Haley, van de Kouwe, Gollub et al., 2006). As a result, it can be hard to determine whether a particular

voxel in one subject is located in the exact same brain region as the same voxel in another subject. And even if they are located in the same brain region, two subjects may activate the same region to differing extents for idiosyncratic reasons. By extension, the presence of activation in one voxel in a sample and the absence of activation in that same voxel in another sample does not necessarily imply a lack of replication across the two samples.

Achieving the scientific ideal of replication would seem hopeless based on this lack of voxels-wise concordance across studies. Fortunately, most social neuroscientists are not concerned about voxel-level replication. Instead, our interest is in understanding the brain at the level of functional regions – the chunks of gray matter that are consistently and coherently involved in a given psychological process. Neighboring voxels, particularly those within the same anatomical structures, tend to perform a similar function, so exact voxel-wise replication is not as important as functional region-wise replication (which is another reason to praise “tedious neuroanatomy”). But this just punts the problem up a level of analysis: If exact voxel-wise replication is not possible, then exact region-wise replication might be challenging as well. How can researchers know which task-related activations replicate and which do not, even at the level of the cluster instead of the voxel?

The answer to this question is highly dependent on how activation is defined across studies. As we have described previously, researchers make a large number of decisions in the course of preprocessing and analyzing their data that could potentially influence the precise size and location of their observed clusters. There are several existing tools to measure the amount of overlap in activation across studies, ranging from simple voxel counting methods (Cohen & DuBois, 1999) to more formal statistical methods (Nichols, Brett, Andersson, Wager, & Poline, 2005), and all of these must be interpreted with the following idea in mind. Even in the ideal case in which the *mental processes* are replicated exactly in two studies, the amount of replication in the *observed data* will vary as a function of fit in design, acquisition, and analysis. These features include, but are by no means limited to, design factors such as the level of complexity and abstraction of the task (see Point #5 later; Plichta, Schwarz, Grimm, Morgen, Mier, Haddad et al., 2012), whether the design is event-related or blocked (Bennett & Miller, 2010), and the subject population (Bosnell, Wegner, Kincses, Kortewek, Agosta, Ciccarelli et al., 2008); acquisition factors including the field strength (Hoenig, Kuhl, & Scheef, 2005), scanning parameters (Bandettini, Wong, Jesmanowicz, Hinks, & Hyde, 1994), and thermal noise (Bodurka, Ye, Petridou, Murphy, & Bandettini, 2007); and analysis factors,

mainly the thresholding procedure (Bennett, Wolford, & Miller, 2009), but also the hemodynamic basis functions (Lindquist & Wager, 2007) and the smoothing kernel (Mikl, Mareček, Hluštík, Pavlicová, Drastich, Chlebus et al., 2008).

Given the complexity of the dependent measure and the chain of decisions that must be made between when participants complete the task and when the data are reported, it seems amazing that fMRI results ever replicate at all. Nonetheless, using the intraclass correlation coefficient as the metric of replication (Caceres, Hall, Zelaya, Williams, & Mehta, 2009), several studies have observed reliabilities of 0.9 or greater (Aron, Gluck, & Poldrack, 2006; Raemaekers, Vink, Zandbelt, van Wezel, Kahn, & Ramsey, 2007). Intraindividual reliability tends to be quite high (Miller, Donovan, Van Horn, German, Sokol-Hessner, & Wolford, 2009), and the majority of variance in fMRI studies has been shown to be from between-subjects variance (Costafreda, Brammer, Vêncio, Mourão, Portela, de Castro et al., 2007). It follows that if variation in between-subjects factors (e.g., early life stress; Taylor, Eisenberger, Saxbe, Lehman, & Lieberman, 2006) is minimized, neural activations may replicate quite faithfully at the level of functional regions. Indeed, there are now several meta-analyses in social and affective neuroscience that point to areas reliably identified for processes of interest such as emotion (Kober, Feldman Barrett, Joseph, Bliss-Moreau, Lindquist, & Wager, 2008), mentalizing (Spreng, Mar, & Kim, 2009; Van Overwalle, 2009), and social cognition more generally (Amodio & Frith, 2006). Further approaches to accounting for between-subject variability are described in Point #5.

5. Why Is There Greater Fundamental Variability in Social Neuroscience Data?

It has been noted in the previously cited reviews (e.g., Bennett & Miller, 2010) that some of the tasks used in social neuroscience tend to have high within-and between-subject variability. In general, we agree and would even conjecture that the source of this variability is that the mental processes studied in social neuroscience – and not just the tasks – are inherently more variable than those studied in other cognitive neurosciences. One reason for this is that social neuroscience tasks often involve manipulations of abstract concepts (e.g., one's own long-term goals or the emotional reactions of other individuals), so people are likely to vary from one another and across time in the strategies they take to engage in the tasks. We refer to these ideas as the *anteriorization-abstraction hypothesis*: that the axis that runs from the posterior to the anterior parts of the

brain (i.e., from the back of the brain to the front) reflects increasing levels of abstraction of mental representation, and that activation changes from focal to diffuse in reliability along that axis. This diffuse activation for abstract representation reflects the high level of cognitive flexibility involved in that process. For example, viewing a picture of a given person relative to considering the mental state of that person can be expected to produce a more posterior/caudal pattern of activation that is more reliable over time and across people.

The anteriorization-abstraction hypothesis suggests some challenges for social neuroscience. The main challenge is that statistics that involve central tendency (e.g., means) are less powerful toward the abstract/anterior end of the gradient. And naturally, our standard statistical tools that are based on the GLM depend on high consistency within subjects (at the first level) and also between subjects (at the second level) to detect effects. Thus, more trials and more subjects are often needed to obtain effects compared to more “reliable” processes. Also, the aim of localization of functions (even abstract ones) can be challenging to meet when, by definition, some functions are more diffuse in their localization than others. What does it mean that top-down control is instantiated in the prefrontal cortex, for instance, if every person uses a slightly different part of the cortex to engage in control and a given person might use a different part at different times? The answer to this question is related to how to define replication in social neuroscience studies (see Point #4), and also, in part, begins to undermine the granularization of function in more anterior parts of the brain. It may be the case that the prefrontal cortex is organized into larger functional units than more posterior regions are, and that any small piece of it has no consistent, specific role beyond top-down control of any or all bottom-up systems (e.g., Heatherton & Wagner, 2011; Miller & Cohen, 2001; Munakata et al., 2011).

So as not to sound too grim, there is also a large upshot that follows from the anteriorization-abstraction hypothesis: more variance means more variance to be explained in a statistical sense. At the within-subjects level, this means explaining how neural processes change across time, for example as a function of changing cognitive strategies over time (Kross, Davidson, Weber, & Ochsner, 2009), or engaging in a process for a sustained versus a brief period (e.g., Somerville, Wagner, Wig, Moran, Whalen, & Kelley, 2013). At the between-subjects level, this means identifying factors that moderate the variation in brain activity across people. This is one reason why subject-level correlations are so appealing: They open a window to explain neural activation using idiographic differences as an alternative to mean tendency. In the future, we

anticipate that more sophisticated multilevel modeling of fMRI data will enhance our ability to explain within-and between-subjects variance, and even the cross-level interactions between the two.

An issue related to increased variance in social neuroscience processes is the notion that the tasks used in social neuroscience are often not *process pure*, meaning that they are comprised of a number of mental processes that cannot be uniquely isolated from one another in terms of their neural underpinnings. For instance, some studies analyze epochs of six or more seconds of thinking in a focused way (e.g., emotion regulation or mindfulness meditation), and experimenters simply have little control over exactly what subjects do and when they do it. Even with the most careful instructions and manipulation checks, we can never know for sure exactly what was happening inside the participants' heads or whether all of them were doing the same thing. (Additionally, even if they honestly told us what they believed they were doing, we know that many mental processes occur outside of awareness.) Experimenters can use *process analysis* (or *cognitive ontology* ; Bilder, Sabb, Parker, Kalar, Chu, Fox et al., 2009) to carefully specify the component operations that are required by a task during the study design phase, and then leverage existing databases (e.g., The Cognitive Atlas; Poldrack et al., 2011) during the analysis phase to triangulate the mental process – neural activation mappings in their data.

6. The Rhetorical Power of Neuroimaging Data

We hope that the main reason for excitement about neuroimaging among lay readers and members of the media is the same as it is for members of the scientific community: that neuroimaging data can provide unique information about the brain systems involved in mental processes and provide insight into the nature of the mental processes themselves. But casual observers have posited that another reason is the compelling nature of the visuals themselves, and data now support this view. For example, one study found that undergraduates rated the scientific reasoning of articles to be more sound when the data were overlaid on a high-resolution brain image compared to the *exact same data* overlaid on a topographical map (McCabe & Castel, 2008). Another study found that even undergraduates with training in neuroscience (but not neuroscience experts) found scientific results more satisfying when they were accompanied by irrelevant neuroscience explanations (Skolnick Weisberg, Keil, Goodstein, Rawson, & Gray, 2008). This result also held for logically flawed arguments, suggesting that the presence of irrelevant neuroscience information can trump

basic reasoning in the minds of lay readers. The authors of this latter study conjectured that the “seductive allure” of neuroscience studies is attributable in part to a cognitive bias toward reductive arguments of mental phenomena. Other factors could include a “technical language” bias (Shafir, Smith, & Osherson, 1990), a “seductive details” effect (Harp & Mayer, 1998), and a “placebic information” heuristic (Langer, Blank, & Chanowitz, 1978).

Whatever their cause, the seemingly mind-numbing effects of neuroscience data among nonexperts should cause concern for those in the field. We highlight three main lessons for researchers that follow from the “seductive allure” effect. First, we need to do a better job educating our students about when neuroimaging data do and do not inform theory. The fact that Ivy League undergraduates who completed an intermediate-level neuroscience class could not differentiate when neuroimaging data were relevant and when they were not is a dramatic illustration of this need (Skolnick Weisberg et al., 2008). Second, we must be exceedingly clear when speaking to members of the media about what the neuroimaging data show. It is difficult enough to get the media to report any scientific results with fidelity, and we have good evidence now that this is even more the case with neuroimaging data. And third, in light of the finding that good scientific explanations with irrelevant neuroimaging data are perceived as *less* satisfying by experts (Skolnick Weisberg et al., 2008), we should be careful not to oversell these data in professional venues.

7. Brain as Predictor: Correlation versus Prediction

Traditional functional neuroimaging studies are designed and analyzed with mental process as the independent variable and brain activation as the dependent variable. In the language of regression, the brain data are the criterion (Y vector) and the task conditions are the predictors (X matrix). This statistical model (i.e., forward inference) has proven to be enormously useful over the past decades for developing and refining an extensive body of knowledge on the mapping from mental process to neural activation. Now, scientists across disciplines are eager to build on this corpus to use neuroimaging data to predict real-world outcomes beyond the laboratory, but the traditional design and analysis tools are insufficient for this purpose. This necessitates a model that flips the traditional one on its head by treating neural activation as the *independent* variable predicting outcomes beyond the scanner in a *brain-as-predictor* approach (Figure 7.11; Berkman & Falk, 2013).

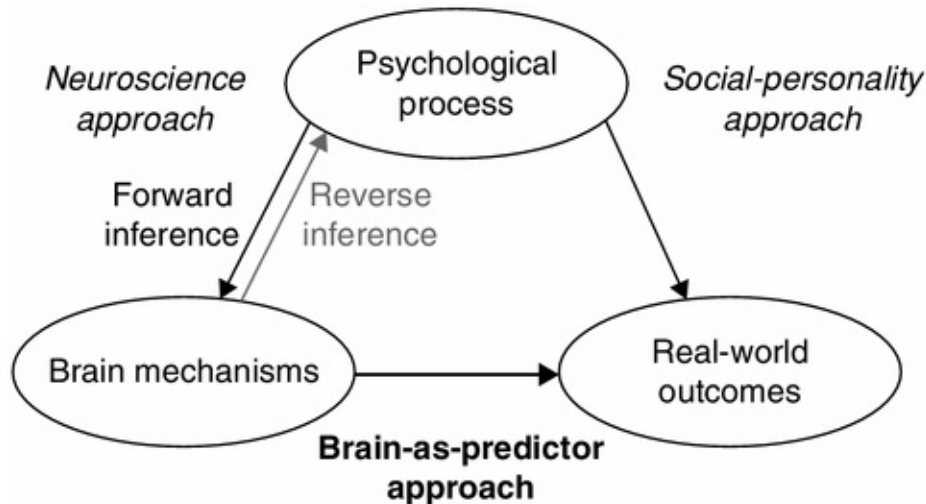


Figure 7.11. The brain-as-predictor approach. Traditionally, social and personality psychologists have been interested in, among other things, mapping the relationship between psychological processes (e.g., cognitions, emotions) and real-world outcomes (e.g., health behavior, discrimination). In contrast, others use neuroimaging tools to map the relationship between psychological process and brain mechanisms. The brain-as-predictor approach integrates these by using brain systems (that have been identified with a specific psychological process) to predict meaningful outcomes beyond the confines of the laboratory.

Traditional correlation approaches in social neuroscience (e.g., those described in Point #2 earlier) are limited because they model the brain as the dependent measure. That is, they answer the question: Which parts of the brain are correlated with an outcome? The design of these studies also commonly measures the outcome concurrently or even before the brain data, which rules out prospective prediction. For example, Mehta and Beer (2010) found that activation in medial orbitofrontal cortex while receiving unfair (versus fair) offers in an ultimatum game was correlated with subsequent aggressive responses. Even though brain activation was measured before the behavioral outcome, the statistical model used was designed to predict neural activity based on the behavior, and not the other way around. This model is highly useful for identifying linear relationships (which is most often its intended use), but it is not a valid tool for prediction when the brain data are the criterion (Gelman & Hill, 2007).

As its name implies, the brain-as-predictor approach models brain activation as the predictor and allows any outcome that occurs after measurement of the activation to be the criterion. For example, we used this approach to test whether

activation in regions hypothesized to be involved in inhibitory control were predictive of self-control in daily life (Berkman, Falk, & Lieberman, 2011). First, in a baseline session, we measured activation in three such regions during a classic inhibitory control task. These were a priori regions selected based on their past involvement in self-control studies. Then, we measured participants' self-control abilities during a cigarette smoking cessation attempt using a daily diary. With all of these data in hand, we entered them simultaneously into a multilevel model, which revealed that increased activation in each of the three brain regions predicted a reduced relationship between cigarette cravings and smoking over the course of three weeks. In another example of this approach, Masten and colleagues used activation in the subgenual anterior cingulate cortex during social rejection to predict subsequent increases in depression among adolescents (Masten, Eisenberger, Borofsky, McNealy, Pfeifer, & Dapretto, 2011).

The brain-as-predictor approach is a useful tool for both theory testing and translational science. In terms of theory, this approach allows for a strong test of the hypothesized function of various brain regions, and can measure not just the presence of predictive validity but also strength. An example of where this might be useful is in testing the specific function of the medial prefrontal cortex, a region thought to be involved in self-processing and comprised of various subdivisions with different hypothesized functions (e.g., Northoff, Qin, & Feinberg, 2011). A researcher interested in a stringent test of these functions could measure the activation in each of the regions and then measure real-world effects of various types of self-processing, for example how new college students select a major (academic self-concept) and how they make new friends (social functioning). The brain-as-predictor approach is also useful in translational settings because, once validated, brain activation could be used to predict clinical outcomes such as response to treatment or substance use relapse. As the cost of imaging goes down and the body of knowledge about brain function goes up, using functional neural activation to predict individual clinical outcomes could be particularly helpful for disorders in which the current diagnostic tools are unavailable, limited, or highly costly.

The largest drawback of this approach is that the hypothesized predicted brain regions must be known in advance. In that sense, brain-as-predictor is not primarily a brain mapping approach; it does not search for an answer to the “where in the brain?” questions (although it can help confirm existing answers to these). Instead, it more directly addresses the “does the brain?” questions, ones that hypothesize a particular function for a region and test whether activation in

that region is a valid predictor of outcomes dependent on that function. To do so, it depends on wholebrain analyses that *do* answer the “where in the brain?” questions for the purpose of hypothesis generation. The brain-as-predictor approach is complementary to traditional forward inference approaches (and distinct from reverse inference). Forward inference and brain-as-predictor are both needed to first identify candidate brain regions involved in a process and then test the predictive validity of those regions outside the scanner (Figure 7.11). We hope that in the future scientists will conduct series of studies that capitalize on the strengths of both methods.

8. Mind Reading?

One of the popular fantasies about neuroimaging is that it can be used to read people's thoughts. Put someone in the scanner or an EEG, the claim goes, and you can know the contents of his mind, including whether one is lying, what one's specific memories look like, and even what one will do in the future, like in the motion picture *Minority Report*. Even though there are several for-profit companies that make these claims, we want to be as clear as possible that this kind of mind reading, which we call *strong mind reading*, does not yet exist, nor is it likely to exist in the foreseeable future. However, there are several kinds of reliable pattern classification that do exist, which we call *weak mind reading*.

Why is strong mind reading so far-fetched? There are dozens of reasons, but we focus on just a few here. First, strong mind reading, if possible, could only occur with a willing participant. Being able to infer a specific mental state requires knowledge of other examples of that state for comparison, often very many of them, and very little useful data would come from someone unwilling to provide them (e.g., a criminal defendant). Second, there is too much variability in how a given mental state (e.g., a feeling of guilt) is represented in the brain both within and between individuals (see Point #5). Even with many examples of a specific mental state to work from (e.g., a memory), the reliability of matching a single new example of that state to previous ones would be quite low. And third, in the case of predicting future actions from current brain states, future behavior is multiply determined, and the extent to which it is determined by information currently in the brain is probably beyond the resolution of current scanners. Experience does change the brain, but these experiences take years to manifest in changes that are visible to our current technology (e.g., Luby, Barch, Belden, Gaffrey, Tillman, Babb, et al., 2012). Even if any single thought, feeling, or memory were visible, we would not even know where to look at this

point.

However, there are a few forms of weak mind reading that may be valid and useful. For instance, we know that decisions are sometimes driven not by consciously accessible thoughts but rather by implicit associations or learning that are not directly accessible to consciousness (Nisbett & Wilson, 1977). Neuroimaging might help explain some variability in decisions or behaviors that would not otherwise be easy to measure. For example, Falk and colleagues found that brain activation in a “neural focus group” predicted changes in population-level behavior (statewide increases in calls to a smoking cessation quitline) in response to health messages (anti-smoking advertisements) above and beyond self-reports about the messages (Falk, Berkman, & Lieberman, 2012). This result suggests that factors that people could not or would not report directly, but were coded in their neural activation patterns, were driving their actual responses to the messages.

Another kind of weak mind reading involves measuring brain activation during many instances of a small number of thoughts (e.g., 100 repetitions each of the nouns “carrot” and “airplane”) and then using pattern classification algorithms to categorize new thoughts (e.g., “helicopter”) into one of the trained categories. A recent study used this technique to demonstrate a 75% accuracy rate on trained words, and about a 60% accuracy on untrained but semantically related words (Mitchell, Shinkareva, Carlson, Chang, Malave, Mason et al., 2008), and another used it to communicate with minimally conscious patients using a yes/no response (Monti, Vanhaudenhuyse, Coleman, Boly, Pickard, Tshibanda et al., 2010). One group of researchers combined pattern classification with real-time analysis to successfully predict decisions in real time during an ultimatum game (Hollmann, Rieger, Baecke, Lützkendorf, Müller, Adolf et al., 2011). As elegant as they are, note that all of these studies either require many exemplars from the same individual or make prediction from one individual to a large group of others (i.e., on average), but none can yield the content of a given thought without at least some prior information.

Conclusion

We have provided a survey of the neuroimaging methods used in social and affective neuroscience, and a discussion of eight of the current controversies and open questions in the field. Our aim was to make this chapter simple but comprehensive so that by its end a reader with no previous neuroimaging

experience would understand the purpose of each of the steps from start to finish of a social neuroscience study, and even have a basic idea of how to begin designing his or her own study. Another aim was to provide thorough explanations of the major ongoing controversies and questions in the field that would be satisfying for those outside it and viewed as even-minded by those inside it. In sum, we hope that this chapter might serve as a comprehensive guide to social neuroscience for our colleagues in social and personality psychology, and also as a healthy introduction to the field for students and faculty seeking to enter it.

In a word, we see the theme of the future of the field as expansion: in theoretical scope, in methodological and statistical depth, in breadth of impact across fields, and in size. A core issue is how to inform social and personality psychology theory using neuroimaging data, and scientists in the field are making encouraging progress in exploring a number of avenues to do so (e.g., brain-as-predictor design and Bayesian meta-analysis). The appeal and accessibility of these methods will increase even more as social neuroscience becomes a standard part of graduate education in social and personality psychology programs. Social and affective neuroscience is also gaining popularity among related fields such as psychopathology, child development, and addiction, and will continue to do so as our methods advance. Finally, developing more powerful statistical and computational methods such as multilevel modeling and structural equation modeling for neuroimaging data will further expand the toolkit available to researchers to test hypotheses with more nuance than null hypothesis significance testing. We hope that this chapter will be a first step into the field for many of those who will contribute to its progress.

Acknowledgments

We are grateful to the UO Social Affective Neuroscience, Developmental Social Neuroscience, and Social Endocrinology Laboratories for helpful feedback on a previous version.

Support for this work was provided by University of Oregon start-up funds to ETB and National Science Foundation (BCS-1122352 & BCS-0819250) to WAC.

Correspondence should be addressed to ETB, Department of Psychology, 1227 University of Oregon, Eugene, OR 97403.

References

- Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, 7, 268–277.
- Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & Covert, A. E. (2004). Neural signals for the detection of unintentional race bias. *Psychological Science*, 15, 88–93.
- Aron, A. R., Gluck, M. A., & Poldrack, R. A. (2006). Long-term test–retest reliability of functional MRI in a classification learning task. *NeuroImage*, 29(3), 1000–1006.
- Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Science*, 8(4), 170–177.
- Bandettini, P. A., Wong, E. C., Jesmanowicz, A., Hinks, R. S., & Hyde, J. S. (1994). Spin-echo and gradient-echo EPI of human brain activation using BOLD contrast: A comparative study at 1.5. *Nuclear Magnetic Resonance in Biomedicine*, 7, 12–20.
- Baumeister, R. F., & Heatherton, T. F. (1996). Self-regulation failure: An overview. *Psychological Inquiry*, 7(1), 1–15.
- Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, 2(4), 396–403.
- Beer, J. S., John, O. P., Scabini, D., & Knight, R. T. (2006). Orbitofrontal cortex and social behavior: Integrating self-monitoring and emotion-cognition interactions. *Journal of Cognitive Neuroscience*, 18(6), 871–879.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1), 289–300.
- Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, 1191(1), 133–155.
- Bennett, C. M., Wolford, G. L., & Miller, M. B. (2009). The principled control of false positives in neuroimaging. *Social Cognitive and Affective Neuroscience*, 4(4), 417–422.

- Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, 8(6), 551–565.
- Berkman, E. T., Burklund, L., & Lieberman, M. D. (2009). Inhibitory spillover: Intentional motor inhibition produces incidental limbic inhibition via right inferior frontal cortex. *NeuroImage*, 47(2), 705–712.
- Berkman, E. T., & Falk, E. B. (2013). Beyond brain mapping: Using neural measures to predict real-world outcomes. *Current Directions in Psychological Science*, 22(1), 45–50.
- Berkman, E. T., Falk, E. B., & Lieberman, M. D. (2011). In the trenches of real-world self-control: Neural correlates of breaking the link between craving and smoking. *Psychological Science*, 22(4), 498–506.
- Berkman, E. T., & Lieberman, M. D. (2010). Approaching the bad and avoiding the good: Lateral prefrontal cortical asymmetry distinguishes between action and valence. *Journal of Cognitive Neuroscience*, 22(9), 1970–1979.
- Berkman, E. T., & Lieberman, M. D. (2011). What's outside the black box?: The status of behavioral outcomes in neuroscience research. *Psychological Inquiry*, 22(2), 100–107.
- Bilder, R. M., Sabb, F. W., Parker, D. S., Kalar, D., Chu, W. W., Fox, J., *et al.* (2009). Cognitive ontologies for neuropsychiatric phenomics research. *Cognitive Neuropsychiatry*, 14(4–5), 419–450.
- Birn, R. M., Cox, R. W., & Bandettini, P. A. (2002). Detection versus estimation in event-related fMRI: Choosing the optimal stimulus timing. *NeuroImage*, 15, 252–264.
- Bodurka, J., Ye, F., Petridou, N., Murphy, K., & Bandettini, P. A. (2007). Mapping the MRI voxel volume in which thermal noise matches physiological noise – Implications for fMRI. *NeuroImage*, 34(2), 542–549.
- Bosnell, R., Wegner, C., Kincses, Z. T., Korteweg, T., Agosta, F., Ciccarelli, O., *et al.* (2008). Reproducibility of fMRI in the clinical setting: Implications for trial designs. *NeuroImage*, 42(2), 603–610.
- Bowman, F. D. B., Guo, Y., & Derado, G. (2007). Statistical approaches to functional neuroimaging data. *Neuroimaging Clinics of North America*, 17(4), 441–458.

- Brant-Zawadzki, M., Gillan, G., & Nitz, W. (1992). MP RAGE: A three-dimensional, T1-weighted, gradient-echo sequence—initial experience in the brain. *Radiology*, 182(3), 769–775.
- Brunet, E., Sarfati, Y., Hardy-Bayle, M. C., & Decety, J. (2000). A PET investigation of attribution of intentions to others with a non-verbal task. *NeuroImage*, 11, 157–166.
- Buckner, R. L., Bandettini, P. A., O’Craven, K. M., Savoy, R. L., Petersen, S. E., Raichle, M. E., *et al.* (1996). Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America*, 93(25), 14878–14883.
- Bug, W. J., Ascoli, G. A., Grethe, J. S., Gupta, A., Fennema-Notestine, C., Laird, A. R., *et al.* (2008). The NIFSTD and BIRNLex vocabularies: Building comprehensive ontologies for neuroscience. *Neuroinformatics*, 6(3), 175–194.
- Bunge, S. A., & Zelazo, P. D. (2006). A brain-based account of the development of rule use in childhood. *Current Directions in Psychological Science*, 15(3), 118–121.
- Buzsaki, G., Traub, R.D., Pedley, T.A. (2003). The cellular basis of EEG activity. In J. S. Ebersole & T. A. Pedley (Eds.), *Current Practice of Clinical Electroencephalography* (3rd ed., pp. 1–11). Philadelphia: Lippincott Williams and Wilkins.
- Caceres, A., Hall, D. L., Zelaya, F. O., Williams, S. C. R., & Mehta, M. A. (2009). Measuring fMRI reliability with the intraclass correlation coefficient. *NeuroImage*, 45(3), 758–768.
- Cacioppo, J. T., & Berntson, G. G. (1992). Social psychological contributions to the decade of the brain: Doctrine of multilevel analysis. *The American Psychologist*, 47(8), 1019–1028.
- Casey, B. J., Cohen, J. D., O’Craven, K., Davidson, R. J., Irwin, W., Nelson, C. A., *et al.* (1998). Reproducibility of fMRI results across four institutions using a spatial working memory task. *NeuroImage*, 8(3), 249–261.
- Cacioppo, J. T., Crites, Jr., S. L., & Gardner, W. L. (1996). Attitudes to the right: Evaluative processing is associated with lateralized late positive event related brain potentials. *Personality and Social Psychology Bulletin*, 22, 1205–1219.

- Cacioppo, J. T., Crites, Jr., S. L., Gardner, W. L., & Berntson, G. G. (1994). Bioelectrical echoes from evaluative categorizations: I. A late positive brain potential that varies as a function of trait negativity and extremity. *Journal of Personality and Social Psychology*, 67, 115–125.
- Chumbley, J. R., & Friston, K. J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage*, 44(1), 62–70.
- Coan, J. A., Schaefer, H. S., & Davidson, R. J. (2006). Lending a hand: Social regulation of the neural response to threat. *Psychological Science*, 17(12), 1032–1039.
- Cohen, M. S., & DuBois, R. M. (1999). Stability, repeatability, and the expression of signal magnitude in functional magnetic resonance imaging. *Journal of Magnetic Resonance Imaging*, 10, 33–40.
- Cools, R., Clark, L., Owen, A. M., & Robbins, T. W. (2002). Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging. *The Journal of Neuroscience*, 22(11), 4563–4567.
- Costafreda, S. G., Brammer, M. J., Vêncio, R. Z. N., Mourão, M. L., Portela, L. A. P., de Castro, C. C., *et al.* (2007). Multisite fMRI reproducibility of a motor task using identical MR systems. *Journal of Magnetic Resonance Imaging*, 26(4), 1122–1126.
- Crites, Jr., S. L., & Cacioppo, J. T. (1996). Electrocortical differentiation of evaluative and nonevaluative categorizations. *Psychological Science*, 7, 318–321.
- Cunningham, W.A. (2010). In defense of brain mapping in social and affective neuroscience. *Social Cognition*, 28(6), 717–722.
- Cunningham, W. A., & Zelazo, P. D. (2007). Attitudes and evaluations: a social cognitive neuroscience perspective. *Trends in Cognitive Sciences*, 11(3), 97–104.
- Dale, A. M., & Buckner, R. L. (1997). Selective averaging of rapidly presented individual trials using fMRI. *Human Brain Mapping*, 5, 329–340.
- Davidson, R. J. (1988). EEG measures of cerebral asymmetry: Conceptual and methodological issues. *International Journal of Neuroscience*, 29, 71–89.

- Davis, J. I., Senghas, A., Brandt, F., & Ochsner, K. N. (2010). The effects of BOTOX injections on emotional experience. *Emotion*, 10(3), 433–440.
- Desmond, J. E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. *Journal of Neuroscience Methods*, 118(2), 115–128.
- Devine, P. G. (1989). Prejudice and stereotypes: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56, 5–18.
- Devlin, J. T., & Poldrack, R. A. (2007). In praise of tedious anatomy. *NeuroImage*, 37, 1033–1041.
- DeWall, C. N., MacDonald, G., Webster, G. D., Masten, C. L., Baumeister, R. F., Powell, C., Combs, D., Schurtz, D. R., Stillman, T. F., Tice, D. M., & Eisenberger, N. I. (2010). Acetaminophen reduces social pain: Behavioral and neural evidence. *Psychological Science*, 21, 931–937.
- Dickenson, J., Berkman, E. T., Arch, J., & Lieberman, M. D. (2013). Neural correlates of focused attention during a brief mindfulness induction. *Social Cognitive and Affective Neuroscience*, 8(1), 40–47.
- Eisenberger, N. I. (2012). The pain of social disconnection: Examining the shared neural underpinnings of physical and social pain. *Nature Reviews Neuroscience*, 13, 421–434.
- Eisenberger, N. I., Berkman, E. T., Inagaki, T. K., Rameson, L. T., Mashal, N. M., & Irwin, M. R. (2010). Inflammation-induced anhedonia: Endotoxin reduces ventral striatum responses to reward. *Biological Psychiatry*, 68, 748–754.
- Evans, A., Kamber, M., Collins, D., & MacDonald, D. (1994). An MRI-based probabilistic atlas of neuroanatomy. In S. Shorvon, D. Fish, F. Andermann, G. Bydder, & H. Stefan (Eds.), *Magnetic Resonance Scanning and Epilepsy* (NATO ASI Series A, Life Sciences ed., Vol. 264, pp. 263–274). New York: Plenum.
- Falk, E. B., Berkman, E. T., & Lieberman, M. D. (2012). From neural responses to population behavior: Neural focus group predicts population level media effects. *Psychological Science*, 23, 439–445.
- Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., *et al.* (2002). Whole brain segmentation: Automated labeling of neuroanatomical

- structures in the human brain. *Neuron*, 33(3), 341–355.
- Foerde, K., Knowlton, B. J., & Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences USA*, 103, 11778–11783.
- Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., Collins, D. L., & the Brain Development Cooperative Group. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54, 313–327.
- Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magnetic Resonance in Medicine*, 33(5), 636–647.
- Friedman, L., Stern, H., Brown, G. G., Mathalon, D. H., Turner, J., Glover, G. H., *et al.* (2008). Test-retest and between-site reliability in a multicenter fMRI study. *Human Brain Mapping*, 29, 958–972.
- Friston, K. (2009). Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLoS Biology*, 7(2), e1000033.
- Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage*, 6(3), 218–229.
- Frith, U., & Frith, C. (2001). The biological basis of social interaction. *Current Directions in Psychological Science*, 10(5), 151–155.
- Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of ‘theory of mind’. *Trends in Cognitive Sciences*, 7(2), 77–83.
- Galván, A., Poldrack, R. A., Baker, C. M., Mcglennen, K. M., & London, E. D. (2011). Neural correlates of response inhibition and cigarette smoking in late adolescence. *Neuropsychopharmacology*, 36(5), 970–978.
- Gehring, W. J., & Fencsik, D. E. (2001). Functions of the medial frontal cortex in the processing of conflict and errors. *The Journal of Neuroscience*, 21(23), 9430–9437.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

- Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15(4), 870–878.
- Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9(4), 416–429.
- Goense, J. B. M., & Logothetis, N. K. (2008). Neurophysiology of the BOLD fMRI signal in awake monkeys. *Current Biology*, 18(9), 631–640.
- Grezes, J., & Decety, J. (2001). Functional anatomy of execution, mental simulation, observation, and verb generation of actions: A meta-analysis. *Human Brain Mapping*, 12(1), 1–19.
- Hallett, M. (2007). Transcranial magnetic stimulation: A primer. *Neuron*, 55(2), 187–199.
- Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science*, 324(5927), 646–648.
- Harmon-Jones, E., & Beer, J. S. (2009). *Methods in Social Neuroscience*. New York: The Guilford Press.
- Harmon-Jones, E., Lueck, L., Fearn, M., & Harmon-Jones, C. (2006). The effect of personal relevance and approach-related action expectation on relative left frontal cortical activity. *Psychological Science*, 17, 434–440.
- Harmon-Jones, E., & Peterson, C. K. (2009). Supine body position reduces neural response to anger evocation. *Psychological Science*, 20(10), 1209–1210.
- Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A cognitive theory of interest in science learning. *Journal of Educational Psychology*, 90(3), 414–434.
- Heatherton, T. F., & Wagner, D. D. (2011). Cognitive neuroscience of self-regulation failure. *Trends in Cognitive Sciences*, 15(3), 132–139.
- Hoenig, K., Kuhl, C. K., & Scheef, L. (2005). Functional 3.0-T MR assessment of higher cognitive function: Are there advantages over 1.5-T Imaging? *Radiology*, 234(3), 860–868.
- Hollmann, M., Rieger, J. W., Baecke, S., Lützkendorf, R., Müller, C., Adolf, D.,

- & Bernarding, J. (2011). Predicting decisions in human social interactions using real-time fMRI and pattern classification. *PLoS ONE*, 6(10), e25304.
- Huettel, S. A., Song, A. W., & McCarthy, G. (2009). *Functional magnetic resonance imaging* (2nd ed.). Sunderland, MA: Sinauer Associates.
- Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., & Turner, R. (2002). Image distortion correction in fMRI: A quantitative evaluation. *NeuroImage*, 16(1), 217–240.
- Ito, T. A., & Cacioppo, J. T. (2000). Electrophysiological evidence of implicit and explicit categorization processes. *Journal of Experimental Social Psychology*, 36, 660–676.
- Ito, T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology*, 75, 887–900.
- Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron*, 58, 284–295.
- Jackson, D. C., Mueller, C. J., Dolski, I., Dalton, K. M., Nitschke, J. B., Urry, H. L. *et al.* (2003). Now you feel it, now you don't: Frontal EEG asymmetry and individual differences in emotion regulation. *Psychological Science*, 75, 887–900.
- Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., *et al.* (2006). Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. *NeuroImage*, 30(2), 436–443.
- Joyce, C. A., & Rossion, B. (2005). The face-sensitive N170 and VPP components manifest the same brain processes: The effect of reference electrode site. *Clinical Neurophysiology*, 116, 2613–2631.
- Kang, Y., Williams, L. E., Clark, M., Gray, J. R., & Bargh, J. A. (2011). Physical temperature effects on trust behavior: The role of insula. *Social Cognitive and Affective Neuroscience*, 6(4), 507–515.
- Kennerley, A. J., Berwick, J., Martindale, J., Johnston, D., Papadakis, N., & Mayhew, J. E. (2005). Concurrent fMRI and optical measures for the investigation of the hemodynamic response function. *Magnetic Resonance in*

Medicine, 54(2), 354–365.

- Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience*, 21(16), RC159.
- Kober, H., Barrett, L. F., Joseph, J., Bliss-Moreau, E., Lindquist, K., & Wager, T. D. (2008). Functional grouping and cortical-subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *NeuroImage*, 42(2), 998–1031.
- Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, 12(5), 535–540.
- Kross, E., Davidson, M., Weber, J., & Ochsner, K. (2009). Coping with emotions past: The neural bases of regulating affect associated with negative autobiographical memories. *Biological Psychiatry*, 65(5), 361–366.
- Kurth, F., Zilles, K., Fox, P. T., Laird, A. R., & Eickhoff, S. B. (2010). A link between the systems: functional differentiation and integration within the human insula revealed by meta-analysis. *Brain Structure and Function*, 214(5–6), 519–534.
- Langer, E., Blank, A., & Chanowitz, B. (1978). The mindlessness of ostensibly thoughtful action: The role of “placebic” information in interpersonal interaction. *Journal of Personality and Social Psychology*, 36(6), 635–642.
- LeDoux, J. A. (1996). *The emotional brain*. New York: Simon & Schuster.
- LeDoux, J. E., Iwata, J., Cicchetti, P., & Reis, D. J. (1988). Different projections of the central amygdaloid nucleus mediate autonomic and behavioral correlates of conditioned fear. *Journal of Neuroscience*, 8, 2517–2529.
- Lieberman, M. D. (2010). Social cognitive neuroscience. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology* (5th ed., pp. 143–193). New York: McGraw-Hill.
- Lieberman, M.D. (2012). A geographical history of social cognitive neuroscience. *NeuroImage*, 61, 432–436.
- Lieberman, M.D., Berkman, E. T., & Wager, T. (2009). Correlations in social neuroscience aren't voodoo: Commentary on Vul *et al.* (2009). *Perspectives*

on *Psychological Science*, 4(3), 299–307.

- Lieberman, M. D., & Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: Re-balancing the scale. *Social Cognitive and Affective Neuroscience*, 4(4), 423–428.
- Lindquist, M. A., & Wager, T. D. (2007). Validity and power in hemodynamic response modeling: a comparison study and a new approach. *Human Brain Mapping*, 28(8), 764–784.
- Luby, J. L., Barch, D. M., Belden, A., Gaffrey, M. S., Tillman, R., Babb, C., *et al.* (2012). Maternal support in early childhood predicts larger hippocampal volumes at school age. *Proceedings of the National Academy of Sciences*, 109(8), 2854–2859.
- Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.
- Masten, C. L., Eisenberger, N. I., Borofsky, L. A., Mcnealy, K., Pfeifer, J. H., & Dapretto, M. (2011). Subgenual anterior cingulate responses to peer rejection: A marker of adolescents' risk for depression. *Development and Psychopathology*, 23(1), 283–292.
- Masten, C. L., Telzer, E. H., & Eisenberger, N. I. (2011). An fMRI investigation of attributing negative social treatment to racial discrimination. *Journal of Cognitive Neuroscience*, 23(5), 1042–1051.
- McCabe, D. P., & Castel, A. D. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition*, 107(1), 343–352.
- McRae, K., Hughes, B., Chopra, S., Gabrieli, J. D. E., Gross, J. J., & Ochsner, K. N. (2010). The neural bases of distraction and reappraisal. *Journal of Cognitive Neuroscience*, 22(2), 248–262.
- Mechelli, A., Price, C. J., Friston, K. J., & Ashburner, J. (2005). Voxel-based morphometry of the human brain: Methods and applications. *Current Medical Imaging Reviews*, 1(2), 105–113.
- Mehta, P. H., & Beer, J. (2010). Neural mechanisms of the testosterone-aggression relation: The role of orbitofrontal cortex. *Journal of Cognitive Neuroscience*, 22(10), 2357–2368.
- Mikl, M., Mareček, R., Hlušík, P., Pavlicová, M., Drastich, A., Chlebus, P., *et*

- al.* (2008). Effects of spatial smoothing on fMRI group inferences. *Magnetic Resonance Imaging*, 26(4), 490–503.
- Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24, 167–202.
- Miller, M. B., Donovan, C.-L., Van Horn, J. D., German, E., Sokol-Hessner, P., & Wolford, G. L. (2009). Unique and persistent individual patterns of brain activity across different memory retrieval tasks. *NeuroImage*, 48(3), 625–635.
- Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2004). Encoding-specific effects of social cognition on the neural correlates of subsequent memory. *Journal of Neuroscience*, 24, 4912–4917.
- Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., *et al.* (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880), 1191–1195.
- Monti, M. M. (2011). Statistical analysis of fMRI time-series: A critical review of the GLM approach. *Frontiers in Human Neuroscience*, 5(28), 1–13.
- Monti, M. M., Vanhaudenhuyse, A., Coleman, M. R., Boly, M., Pickard, J. D., Tshibanda, L., *et al.* (2010). Willful modulation of brain activity in disorders of consciousness. *The New England Journal of Medicine*, 362(7), 579–589.
- Moosman, M., Eichele, T., Nordby, H., Hugdahl, K., & Calhoun, V. D. (2008). Joint independent component analysis for simultaneous EEG–fMRI: Principle and simulation. *International Journal of Psychophysiology*, 67, 212–221.
- Mumford, J., & Nichols, T. E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *NeuroImage*, 39, 261–268.
- Munakata, Y., Herd, S. A., Chatham, C. H., Depue, B. E., Banich, M. T., & O'Reilly, R. C. (2011). A unified framework for inhibitory control. *Trends in Cognitive Sciences*, 15(10), 453–459.
- Murphy, K., & Garavan, H. (2004). An empirical investigation into the number of subjects required for an event-related fMRI study. *NeuroImage*, 22, 879–885.
- Nichols, T., Brett, M., Andersson, J., Wager, T., & Poline, J.-B. (2005). Valid conjunction inference with the minimum statistic. *NeuroImage*, 25(3), 653–

660.

- Nichols, T., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, 12(5), 419–446.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259.
- Northoff, G., Qin, P., & Feinberg, T. E. (2011). Brain imaging of the self – Conceptual, anatomical and methodological issues. *Consciousness and Cognition*, 20(1), 52–63.
- Ochsner, K. N., Bunge, S. A., Gross, J. J., & Gabrieli, J. D. E. (2002). Rethinking feelings: An fMRI study of the cognitive regulation of emotion. *Journal of Cognitive Neuroscience*, 14(8), 1215–1229.
- Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The enigmatic temporal pole: A review of findings on social and emotional processing. *Brain*, 130(7), 1718–1731.
- Pascual-Marqui, R. D., Michel, C. M., & Lehmann, D. (1994). Low resolution electromagnetic tomography: A new method for localizing electrical activity in the brain. *International Journal of Psychophysiology*, 18, 49–65.
- Pfeifer, J. H., Masten, C. L., Borofsky, L. A., Dapretto, M., Fuligni, A. J., & Lieberman, M. D. (2009). Neural correlates of direct and reflected self-appraisals in adolescents and adults: When social perspective-taking informs self-perception. *Child Development*, 80(4), 1016–1038.
- Pierce, K. (2011). Early functional brain development in autism and the promise of sleep fMRI. *Brain Research*, 1380(C), 162–174.
- Plichta, M. M., Schwarz, A. J., Grimm, O., Morgen, K., Mier, D., Haddad, L., *et al.* (2012). Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. *NeuroImage*, 60, 1746–1758.
- Plonsey, R. (1963). Reciprocity applied to volume conductors and the EEG. *IEEE Transactions on Biomedical Engineering*, 19, 9–12.
- Poldrack, R. (2011, October 4). The iPhone and the brain. *The New York Times*. Retrieved July 25, 2013 from <http://www.nytimes.com/2011/10/05/opinion/theiphone-and-the-brain.html>.

- Poldrack, R., Fletcher, P., Henson, R., Worsley, K., Brett, M., & Nichols, T. (2008). Guidelines for reporting an fMRI study. *NeuroImage*, 40(2), 409–414.
- Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, 10(2), 59–63.
- Poldrack, R. A. (2010). Mapping mental function to brain structure: How can cognitive neuroimaging succeed? *Perspectives on Psychological Science*, 5(6), 753–761.
- Poldrack, R. A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y. *et al.* (2011). The cognitive atlas: Toward a knowledge foundation for cognitive neuroscience. *Frontiers in Neuroinformatics*, 5, 1–11.
- Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. New York: Cambridge University Press.
- Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R. J. A., Kahn, R. S., & Ramsey, N. F. (2007). Test-retest reliability of fMRI activation during prosaccades and antisaccades. *NeuroImage*, 36(3), 532–542.
- Redcay, E., Dodell-Feder, D., Pearrow, M. J., Mavros, P. L., Kleiner, M., Gabrieli, J. D. E., & Saxe, R. (2010). Live face-to-face interaction during fMRI: A new tool for social cognitive neuroscience. *NeuroImage*, 50, 1639–1647.
- Redcay, E., Kennedy, D. P., & Courchesne, E. (2007). fMRI during natural sleep as a method to study brain function during early childhood. *NeuroImage*, 38(4), 696–707.
- Rugg, M. D., Schloerscheidt, A. M., Doyle, M. C., Cox, C. J., & Patching, G. R. (1996). Event-related potentials and the recollection of associative information. *Cognitive Brain Research*, 4, 297–304.
- Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, 300(5626), 1755–1758.
- Sareen, J., Campbell, D. W., Leslie, W. D., Malisza, K. L., Stein, M. B., Paulus, M. P. *et al.* (2007). Striatal function in generalized social phobia: A functional magnetic resonance imaging study. *Biological Psychiatry*, 61(3), 396–404.
- Saxe, R., & Powell, L. (2006). It's the thought that counts: Specific brain regions

- for one component of theory of mind. *Psychological Science*, 17(8), 692–699.
- Scherg, M., & Ebersole, J.S. (1993). Models of brain sources. *Brain Topography*, 5, 419–423.
- Shafir, E. B., Smith, E. E., & Osherson, D. N. (1990). Typicality and reasoning fallacies. *Memory & Cognition*, 18(3), 229–239.
- Shattuck, D. W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K. L. *et al.* (2008). Construction of a 3D probabilistic atlas of human cortical structures. *NeuroImage*, 39(3), 1064–1080.
- Skolnick Weisberg, D., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, 20(3), 470–477.
- Small, D. M., Gerber, J. C., Mak, Y. E., & Hummel, T. (2005). Differential neural responses evoked by orthonasal versus retronasal odorant perception in humans. *Neuron*, 47(4), 593–605.
- Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., *et al.* (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23, S208–S219.
- Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, 44(1), 83–98.
- Somerville, L. H., Wagner, D. D., Wig, G. S., Moran, J. M., Whalen, P. J., & Kelley, W. M. (2013). Interactions between transient and sustained neural signals support the generation and regulation of anxious emotion. *Cerebral Cortex*, 23(1), 49–60.
- Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *Journal of Cognitive Neuroscience*, 21(3), 489–510.
- Spunt, R. P., & Lieberman, M. D. (2012). Dissociating modality-specific and supramodal neural systems for action understanding. *Journal of Neuroscience*, 32(10), 3575–3583.
- Stice, E., Yokum, S., Burger, K. S., Epstein, L. H., & Small, D. M. (2011).

Youth at risk for obesity show greater activation of striatal and somatosensory regions to food. *Journal of Neuroscience*, 31(12), 4360–4366.

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain*. New York: Thieme.

Taylor, S. E., Eisenberger, N. I., Saxbe, D., Lehman, B. J., & Lieberman, M. D. (2006). Neural responses to emotional stimuli are associated with childhood family stress. *Biological Psychiatry*, 60(3), 296–301.

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811), 515–518.

Tomarken, A. J., Davidson, R. J., Wheeler, R. E., & Doss, R. C. (1992). Individual differences in anterior brain asymmetry and fundamental dimensions of emotion. *Journal of Personality and Social Psychology*, 62, 676–687.

Tulving, E., Schacter, D. L., & Stark, H. A. (1982). Priming effects in word-fragment completion are independent of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(3), 336–342.

Uddin, L. Q., Clare Kelly, A. M., Biswal, B. B., Xavier Castellanos, F., & Milham, M. P. (2009). Functional connectivity of default mode network components: Correlation, anticorrelation, and causality. *Human Brain Mapping*, 30(2), 625–637.

Urry, H. L., Nitschke, J. B., Dolski, I., Jackson, D. C., Dalton, K. M., Mueller, C. J., Rosenkranz, M. A., Ryff, C. D., Singer, B. H., & Davidson, R. J., (2004). Making a life worth living: Neural correlates of well-being. *Psychological Science*, 15, 367–372.

Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Human Brain Mapping*, 30(3), 829–858.

van Veen, V., & Carter, C. S. (2002). The timing of action-monitoring processes in the anterior cingulate cortex. *Journal of Cognitive Neuroscience*, 14, 593–602.

Viswanathan, A., & Freeman, R. D. (2007). Neurometabolic coupling in cerebral cortex reflects synaptic more than spiking activity. *Nature Neuroscience*, 10(10), 1308–1312.

- Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, 4(3), 274–290.
- Wager, T. D., & Nichols, T. E. (2003). Optimization of experimental design in fMRI: a general framework using a genetic algorithm. *NeuroImage*, 18(2), 293–309.
- Waytz, A., & Mitchell, J. P. (2011). Two mechanisms for simulating other minds: Dissociations between mirroring and self-projection. *Current Directions in Psychological Science*, 20(3), 197–200.
- Weiskopf, N., Hutton, C., Josephs, O., Turner, R., & Deichmann, R. (2007). Optimized EPI for fMRI studies of the orbitofrontal cortex: compensation of susceptibility-induced gradients in the readout direction. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 20(1), 39–49.
- Wheatley, T., Milleville, S. C., & Martin, A. (2007). Understanding animate agents: Distinct roles for the social network and mirror systems. *Psychological Science*, 18, 469–474.
- Williams, K. D., Cheung, C. K., & Choi, W. (2000). Cyberostracism: Effects of being ignored over the internet. *Journal of Personality and Social Psychology*, 79(5), 748–762.
- Wilms, M., Schilbach, L., Pfeiffer, U., Bente, G., Fink, G. R., & Vogeley, K. (2010). It's in your eyes – using gaze-contingent stimuli to create truly interactive paradigms for social cognitive and affective neuroscience. *Social Cognitive and Affective Neuroscience*, 5(1), 98–107.
- Xu, J., Monterosso, J., Kober, H., Balodis, I. M., & Potenza, M. N. (2011). Perceptual load-dependent neural correlates of distractor interference inhibition. *PLoS ONE*, 6(1), e14552.
- Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8), 665–670.

¹ Of course, it is possible and necessary at times to run a study twice: once as an fMRI study and once as an EEG study. Further, new technologies are being

developed to collect EEG and fMRI data simultaneously, and statistical methods are being developed to fuse the two types of data together (see Moosman, Eichele, Nordby, Hugdahl, & Calhoun, [2008](#)).

Chapter eight Behavior Genetic Research Methods

Testing Quasi-Causal Hypotheses Using Multivariate Twin Data

Eric Turkheimer and K. Paige Harden

I find the use of a correlation coefficient a dangerous symptom. It is an enemy of generalization, a focuser on the “here and now” to the exclusion of the “there and then.” Any influence that exerts selection on one variable and not the other will shift the correlation coefficient. What usually remains constant under such circumstances is one of the regression coefficients. If we wish to seek for constancies, then, regression coefficients are much more likely to serve us than correlation coefficients.

(Tukey, [1969](#), p. 89).

Introduction

The suggested topic of this chapter – methods for behavioral genetic research in personality – is potentially misleading. Behavioral genetics, of course, is simply the science of genetics in its many forms as it is applied to behavior, and for the most part there is no reason for genetic research methods for the study of behavior to be any different than genetic methods applied to nonbehavioral characteristics of organisms. The genetics of behavior can be studied in humans or in nonhuman animals, using correlational or experimental methods; it can be inferred from patterns of familial relations or observed more directly in DNA. Like any characteristic of an organism, behavior can be thought of in terms of individual differences or unvarying species-typical characteristics; it can be studied in the cross-sectional context of the current moment or as an ongoing process in a life span or evolutionary time; and it can be seen as an aspect of normal functioning or as a reflection of disorder and distress. The same is true of personality. There is little about personality that requires it to be studied differently than diabetes, or height, for that matter. The domain of behavioral genetic research methods in personality is in principle no less extensive than the intersection of genetics and personality. As such it is too vast to review in a single chapter.

Research methods in both behavioral genetics and personality are currently at a crossroads. Although the history of the behavioral genetics of personality has its origins in animal breeding, and the foundational work in the field is largely about temperament in dogs (Scott & Fuller, 1965), what has come to be thought of as behavioral genetics are the methods of quantitative genetics, in which genetic and environmental processes are inferred from differences in genetic and environmental relationships in twin and sibling pairs, families, and pedigrees. Over roughly the same period, “personality” has come to refer largely to individual differences in human personality, especially as they are assessed via paper-and-pencil and self-report. The intersection of these more specific paradigms, in which correlations between the self-reported personality scores of family members are analyzed using quantitative genetic statistical models, has defined the behavioral genetics of personality for the last 50 years (Tellegen, Lykken, Bouchard, Wilcox, Segal, & Rich, 1988).

That methodological era is coming to an end. The wider availability of specific genetic markers and the sequencing of the human genome has supplemented, and to some degree supplanted, the classical quantitative methods of the last century (Charney, 2012). Researchers in personality have recognized the limitations of a narrow focus on self-report (Oltmanns & Turkheimer, 2009), and the recent explosion in evolutionary thinking in the behavioral sciences has had a profound effect on the field (Penke, Denissen, & Miller, 2007), in particular by rekindling interest in species-general aspects of personality as opposed to individual differences, and by reminding us of the importance of personality in nonhuman animals (Gosling & John, 1999).

With that in mind, it may seem retrograde to focus a review of behavioral genetic research methods in personality on twin methods for the study of individual differences in self-reported responses in humans. The choice can be defended on several grounds, other than the simple fact that this is where the expertise of the authors happens to lie. First is to counter the too-frequent tendency in the behavioral sciences to move on from one poorly understood method to the next, motivated not by the theoretical completion of the old paradigm but rather by the availability of new technology. This review endeavors to show that both the foundations and implications of classical twin studies of personality have not been fully understood. Related to this motivation, the new molecular genetic methodologies have themselves led to a complex tangle of methodological difficulties, which one of us has recently reviewed elsewhere (Turkheimer, 2012), although not in the context of personality per se

(but see Munafo, Clark, Moore, Payne, Walton, & Flint, 2003). Finally, there is the undeniable but somewhat mysterious fact that notwithstanding the thousands of twin studies of personality that have been conducted, and an equally rich history of theoretical writing on the subject, the quantitative genetics of personality remains stubbornly controversial, both widely accepted as foundational yet regularly rejected as misleading or worse. Its merits and implications continue to be debated in the top journals (Charney, 2012). The perhaps unrealistic goal of this chapter is to soften the disagreement about the genetics of behavior by reformulating its methodological foundation of twin and family studies. Later, we also apply our reformulation of older methods to gain realistic understanding of the newer ones that capitalize on the availability of measured DNA.

Personality as Nonexperimental Science

Focusing the chapter on individual differences in humans highlights a particularly problematic aspect of scientific inference in the human behavioral sciences: the inference of causality from nonexperimental data (see West, Cham, & Liu, Chapter 4 in this volume). It is, of course, possible to study personality experimentally, using random assignment to experimental conditions to isolate causal effects of manipulations from extraneous variables that might otherwise confound them. The branch of personality psychology that interfaces with social psychology consists largely of this kind of work. It is even possible to conduct randomized experimental research while including genetic information (Burt, 2009), although this is not often attempted in humans.

When studying human individual differences in personality, the observations are usually correlational, beginning with the most fundamental observations in personality, the patterns of association among personality items that have been the basis for factor analytic studies of personality structure since Cattell (1957). Even at a more molar level, the basic observations of personality science usually involve statistical associations, either among the personality traits themselves or with external variables that are indicators of validity. We refer to these relations as phenotypic associations, with “phenotype” denoting a characteristic of an organism at the observational level, as opposed to its underlying causes. Phenotypic associations with personality are easy to observe, but in nonexperimental work the important underlying questions are about cause: What causes individual differences in personality, and what do individual differences in personality cause? For example, does neuroticism cause poor physical health

(Shipley, Weiss, Der, Taylor, & Deary, 2007)? Does military service change one's personality, or are men with certain personality characteristics more likely to select military service (Jackson, Thoemmes, Jonkmann, Ludtke, & Trautwein, 2012)? Do changes in impulsivity cause a young adult to “mature out” of alcohol use, or does heavy drinking cause increases in impulsivity (Littlefield, Sher, & Wood, 2009; Quinn, Stappenbeck, & Fromme, 2011)? These causal questions are very difficult to answer, and phenotypic associations alone are causally ambiguous. The only way to demonstrate conclusively that a phenotypic association between heavy drinking and impulsivity is causal would be to assign individuals randomly to different heavy-drinking conditions and see what happens. Random experimentation of this kind is, of course, often impossible for practical or ethical reasons. In the absence of random assignment, how can a social scientist proceed?

As is usually the case in the social sciences, the answer is that social scientists can resort to quasi-experimental methods, in the hope of capturing some of the causal certainty offered by the idealized random experiment (see West et al., Chapter 4 in this volume). Suppose, for example, there existed pairs of children who had been matched for cultural background and genetic predisposition, yet who nonetheless differed in some hypothesized causal factor. To continue with the example of heavy drinking and impulsivity, if there were pairs of children matched for genetic and environmental family background but differing in their drinking habits, and if differences in drinking within pairs of matched children was still related to their personality, this association could not be the result of environmental or genetic family background, because the pairs had been matched for these traits. Associations within matched pairs would not *prove* causation, because matching can never be comprehensive and perfect, but to the extent the matching succeeded in holding constant important confounds, the within-pair associations would strengthen our impression that the association may have a causal basis. We have adopted the qualified term *quasi-causal* to denote associations that have survived analysis using quasi-experimental methods.

Needless to say, the matched pairs we have described do exist: they are called identical (monozygotic; MZ) twins reared together. Other kinds of familial clusters – fraternal (dizygotic; DZ) twins, siblings, half-siblings, twins reared apart, adoptive siblings, and so forth – are also matched, but to a lesser degree than are identical twins reared together. This chapter makes the case that the essential contribution of what is commonly called behavior genetics is the use of such familial clusters to obtain a significant but imperfect degree of quasi-

experimental control over nonexperimental phenotypic associations. To illustrate this point, we begin by presenting a regression-based analysis of MZ twin data on religiosity (involvement in organized religious activities) and delinquent behavior during adolescence. This starting-off point differs from what is typically thought of as a “twin study” in two important respects. First, the focus of our analysis is on the relation between two individual differences variables, rather than on dividing the sources of variation in a single behavior into genetic versus environmental components. By the end of this first analysis, we will not know much about how much genes matter for either religiosity or delinquency, but we will know much more about how they are related to each other. Second, we begin by using data from only MZ twins rather than from both MZ and DZ twins. As we describe later, the familiar decomposition of observed variance into genetic and environmental components depends on comparing the relative similarity of MZ versus DZ twins, but we hope to convince the reader that the clearest exposition of the twin method starts elsewhere. Indeed, we request that the reader lay aside what he or she already knows or has heard about twin studies, including the idea that the purpose of behavioral genetics is to estimate the magnitude of genetic and environmental contributions to a trait. In later sections, we expand on our simple MZ-twin regression analysis to show how it intersects with more complex – and perhaps more familiar – methods for analyzing twin data. (The reader interested in a more traditional introduction to the twin method can see Plomin, DeFries, Knopik, & Neiderhiser, 2012 for an exhaustive account, or Neale & Maes, 2007 for a more computationally oriented approach.).

Religiosity and Delinquency in MZ Twins

To provide a brief substantive background for our example, the incidence of delinquency increases so dramatically in adolescence that some researchers consider it to be developmentally normative (Moffitt, 1993). Adolescents commit more than 30% of major crimes in the United States (Federal Bureau of Investigation, 2004). One potential protective factor against delinquency is religiosity – that is, affiliation with and involvement in religious organizations and activities. Religious involvement may decrease problem behavior by instilling beliefs about divine sanctions, encouraging prosocial ties that foster concern for collective well-being, facilitating the intergenerational communication of conforming values, and buffering against psychological distress that otherwise may be acted out in problem behavior (Alpert, 1939;

Smith, 2003). A meta-analysis of 60 studies concluded that there is a moderate negative relationship between religiosity and delinquency (Baier & Wright, 2001). However, the association between religiosity and delinquency is confounded by numerous variables related to both, including genetic factors (Koenig, McGue, Krueger, & Bouchard, 2005; Miles & Carey, 1997).

Twin data on religiosity and delinquency are drawn from the National Longitudinal Study of Adolescent Health (Add Health), a nationally representative study designed to assess adolescent health and risk behavior, collected in four waves between 1994 and 2008. Add Health participants were recruited using a stratified school-based sampling design. A randomly selected subsample of 20,745 participants (randomly selected from school rosters) completed a 90-minute in-home interview between April and December 1995 (Wave I interview; 10,480 female, 10,264 male). Participants ranged in age from 11 to 21 years ($M = 16$ years, 25th percentile = 14 years, 75th = 17 years). The design features of the Add Health data set have been extensively described elsewhere (Harris, 2011), and interested readers are referred to the Add Health website (<http://www.cpc.unc.edu/addhealth>) for additional information. In this chapter we use a subsample of the Add Health participants that comprises all adolescents between 11 and 20 years old who were identified as monozygotic (MZ) or dizygotic (DZ) twins raised together in the same household ($N = 289$ MZ pairs, 451 DZ pairs). Twin zygosity was determined primarily on the basis of self-report and four questionnaire items concerning how often twins were confused with one another and the similarity of their physical appearance. Eighty-nine pairs of uncertain zygosity were determined to be identical if they shared five or more genetic markers. Details regarding the Add Health twins sample are described by Harris, Halpern, Smolen, and Haberstick (2006).

Religiosity was measured using four items (rated on four-point or five-point ordinal scale) assessing importance of religion, frequency of prayer, attendance at religious services, and attendance at youth groups. Additional religiosity items that focused primarily on type of affiliation (e.g., identification as born-again) or theological beliefs (e.g., divine authorship of sacred texts) were excluded. Individuals who denied any religious affiliation were not assessed for religiosity during the interview; they were assigned scores as appropriate (e.g., “Never” for frequency of prayer; “Not at all important” for importance of religion.) Items scored in reverse direction, such that higher scores reflect less religiosity, were reversed numerically. Religiosity scores were computed by summing responses on the four items ($M = 9.44$, $SD = 4.22$, median = 9, range = 4 to 17, $\alpha = 0.76$).

Adolescents were also asked how often in the last 12 months they had engaged in each of 15 antisocial behaviors: Never (0), One or Two Times (1), Three or Four Times (2), or Five or More Times (3). In addition, they were asked how often in the last 12 months each of 4 violent events happened: Never (0), Once (1), More Than Once (2). A previous confirmatory factor analysis of this data indicated that 11 items pertaining to theft, deception, and public rowdiness were indicative of a single factor, labeled here as delinquency (Harden, Mendle, Hill, Turkheimer, & Emery, 2008). (The remaining items were indicative of a factor pertaining to aggressive or violent behavior.) Delinquency factor scores ($M = 0.12$, $SD = 0.81$, range = -1.11 to 3.58 , 25th percentile = -0.50 , 75th percentile = 0.68) were estimated using the program Mplus (Muthén & Muthén, 1998–2010).

Random Effects Models

The most readily apparent analytic complexity introduced by MZ twin data (as opposed to data on singletons) is that observations may no longer be considered independent: individuals are clustered within twin pairs. Failure to consider this nonindependence may cause serious bias in the estimation of parameters and standard errors. Random effects models, also known as hierarchical linear models or mixed effects models, are a popular approach for the analysis of clustered data. (For a comprehensive introduction to random effects models, see Raudenbush and Bryk (2002) or Schoemann, Rhemtulla, and Little, Chapter 21 in this volume.) A basic mixed effects model for our data, analogous to a simple regression in non-clustered data, is as follows:

$$Y_{ij} = B_{00} + (B_{01} \times X_{ij}) + u_{0j} + e_{ij} \quad (8.1)$$

The subscripts ij represent the i^{th} twin within the j^{th} pair. The first part of Model 1 is directly comparable to traditional regression analysis, with B_{00} representing the population intercept and β_{01} representing the expected increase in delinquency given an increase in one unit religiosity. Together this represents the fixed portion of the model. The latter, random portion of the model is composed of u_{0j} , the pair-level error of prediction (i.e., the difference between the population-level intercept B_{00} and the intercept for a given twin pair); and e_{ij} , the individual-level error of prediction. As applied to our example, u_{0j} reflects the extent to which a twin pair is, on average, less or more delinquent than the

overall population; e_{ij} reflects the extent to which an individual twin is less or more delinquent than the pair average. Unaccounted-for variation in delinquency is thus divided into two components: one part shared by twins in a pair and another part independent among individual twins. Taken together, the fixed effects and random effects parts comprise a model closely related to traditional regression analysis; the only addition is an estimate of the dependence between observations within a cluster (i.e., the pair-level residual variation).

This model was estimated in MZ twin pairs using PROC MIXED in SAS (see Appendix A at the end of the chapter for code). Results are listed under Model 1 in Table 8.1. Consistent with previous epidemiological research, religiosity was significantly associated with lower delinquency ($\beta = 0.027$). Of the total unaccounted-for variation in delinquency ($0.324 + 0.343 = 0.667$), 51.4% ($0.343/0.667$) was attributable to genetic and environmental factors that make twins similar. Put another way, the intraclass correlation for twin pairs was 0.51. The remaining 48.6% of the variance existed within twin pairs and can thus be attributed to environmental influences that make twins different and to measurement error.

Table 8.1. Results from Mixed Effects Models of Religiosity and Delinquency

	MZ Twins Only			MZ and DZ Twins
	Model 1	Model 2	Model 3	Model 4
Fixed Effects				
Intercept	-.178 (.097)	-.353 (.113)*	-.353 (.113)*	-.404 (.069)*
Religiosity	.027 (.009)*	-.008 (.017)		
Religiosity Deviation			-.008 (.017)	-.008 (.018)
Deviation*Zygosity (DZ Effect – MZ Effect)				.027 (.021)
Religiosity Pair Average		.055 (.020)*	.047 (.011)*	.050 (.007)*
Random Effects				
MZ Twin Pair	.343 (.048)*	.335 (.046)*	.335 (.046)*	.314 (.045)*
Within-MZ Residual	.324 (.029)*	.317 (.028)*	.317 (.028)*	.317 (.028)*
DZ Twin Pair				.225 (.031)*
Within-DZ Residual				.389 (.028)*

* Significant at $P < .05$.

Besides dividing unaccounted-for variation in delinquency into between-pair

and within-pair components (and providing accurate estimates of the standard errors), the results of Model 1 tell us nothing that we could not have known from an ordinary correlational study. There is, however, an additional piece of information that we can include in the model – the average level of religiosity for the twin pair (\bar{X}_{0j}):

$$Y_{ij} = B_{00} + (B_{01} \times X_{ij}) + (B_{02} \times \bar{X}_{0j}) + u_{0j} + e_{ij} \quad (8.2)$$

Consider how the addition of this one piece of information changes the meaning of the parameter B_{01} , which now quantifies whether an individual who is more religious is less delinquent, *controlling for the overall level of religiosity in his or her twin pair*. Because religiosity is not randomly assigned, controlling for the average level of religiosity in the pair essentially controls for being from the “type” of family that is religious, including all the between-family genetic and environmental differences that are confounded with average religiosity. Results from this model are listed under Model 2 in [Table 8.1](#). Twin pairs who are more religious, on average, have lower levels of delinquency ($\beta_{02} = 0.055$). However, if Twin A is more religious than Twin B, this within-pair difference does *not* significantly predict delinquency ($\beta_{01} = -0.008$), as would be predicted by a causal hypothesis.

A difficulty with including the pair average as a predictor is that it may be strongly correlated with an individual's score. To ameliorate the problem of multicollinearity, Model 2 may be reparameterized to yield orthogonal covariates, namely the twin-pair average (\bar{X}_{0j}) and the *deviation* of each twin from the twin-pair average ($X_{ij} - \bar{X}_{0j}$):

$$Y_{ij} = B_{00} + (B_B \times \bar{X}_{0j}) + (B_W \times (X_{ij} - \bar{X}_{0j})) + u_{0j} + e_{ij} \quad (8.3)$$

The between-cluster regression coefficient B_B estimates whether pairs with higher average religiosity have lower average delinquency, including the effects of the unmeasured covariates that vary at the pair level. In contrast, the within-cluster regression coefficient B_W estimates whether the MZ twin with higher religiosity than his or her co-twin also has lower delinquency than his or her co-twin. Results from this model (Model 3) are shown in [Table 8.1](#). The regression of delinquency on religiosity within twin pairs is *not* significant ($\beta_W = -0.008$),

but the regression on mean pair religiosity is significant ($\beta_B = 0.047$). Model 3 is merely a reparameterization of Model 2, such that B_B in Model 3 equals the sum of the two regression coefficients from Model 2 ($.053 - .008 = .047$), but because the covariates are orthogonal the standard error for the between-cluster regression is slightly smaller in Model 3.

Again, how to interpret these results? The within-cluster regression is not biased by the exclusion of cluster-level confounds; therefore, the within-cluster regression better approximates the “true” quasi-causal relation between religiosity and delinquency in the population. In other words, B_W is the key parameter of interest for evaluating a quasi-causal hypothesis. In the current analysis, the quasi-causal relation between religiosity and delinquency appears to be nonexistent: Families who are more religious are less delinquent, but a twin who is more religious than his co-twin is not. Another way of making the same point is to observe that to the extent the relationship is ultimately causal, the structural relation between religiosity and delinquency should not differ depending on whether one is comparing means of twin pairs, deviations of individual twins from their pair mean, or unrelated individuals (see the discussion of rat pups in Turkheimer & Waldron, 2000). These comparisons are invariant causally but differ in their confounds: twin comparisons are confounded only by environmental influences unique to each twin, while comparisons between unrelated individuals are confounded by all genetic and environmental differences between families. Inequality of B_W and B_B , therefore, suggests the operation of third-variable confounds operating between families.

Comparison of within-and between-cluster regression coefficients is a strategy found frequently in the medical literature, with the aim of disentangling “maternal” factors from “fetal origins” of disease aetiology (for a review, see Carlin, Gurrin, Sterne, Morley, & Dwyer, 2005). Mixed effects models of twin data have been productively applied to the study of birth weight and cord blood erythropoietin (Morley, Moore, Dwyer, Owens, Umstad, & Carlin, 2005), birth weight and blood pressure (Mann, De Stavola, & Leon, 2004), and tobacco use and bone density (Hopper & Seeman, 1994), to name just a few examples. This statistical method would be just as productively applied to the study of psychological development as to the study of disease.

Differentiating Genetic and Shared Environmental Confounds

Thus far, we have considered average family religiosity and within-twin pair differences in religiosity in pairs of identical twins. The former effect is confounded by all genetic and environmental factors that vary between families and are systematically associated with religiosity. The latter effect is confounded only by those factors that vary within twin pairs.

Inclusion of DZ twin pairs complicates the analysis of between-and within-pair variances to some extent, but ultimately allows the estimation of a third variance, and with the addition of some statistical and biological assumptions leads to the familiar terms of the classical twin model. In MZ twins, who are identical genetically, within-pair confounds are necessarily environmental in origin. In DZ twins, who share only 50% of their genes, there are both genetic and environmental within-pair confounds. Therefore, to the extent that the relation between religiosity and delinquency is attributable to *genetic* confounds, there should be a larger within-pair effect for DZ twin pairs than for MZ twin pairs. To model this, we can specify an additional interaction term to our mixed effects model:

$$\begin{aligned}
 Y_{ij} = & B_{00} + (B_B \times \bar{X}_{0j}) + (B_W \times (X_{ij} - \bar{X}_{0j})) \\
 & + (B_{03} \times ZYG) + (B_{04} \times ZYG \\
 & \times (X_{ij} - \bar{X}_{0j})) + u_{0j} + e_{ij}
 \end{aligned}
 \tag{8.4}$$

When zygosity (abbreviated ZYG) is dummy-coded as 0 in MZ twins and 1 in DZ twins, B_W estimates the within-pair regression for MZ twins, whereas B_{04} estimates the difference between MZ and DZ twins in the within-pair effect. To the extent that confounding variables are genetic in origin, the within-pair effect will be larger for DZ twins than MZ twins. In contrast, MZ and DZ twin pairs raised together control equally well for shared environmental factors, thus there will be no difference between pair types if the relevant confounds are environmental in origin. There is no reason to expect a main effect of zygosity on the outcome of interest; however, it is necessary to include the main effect of a covariate included in an interaction term.

Results from this model (Model 4), using data from both MZ and DZ twin pairs, are shown in [Table 8.1](#). See Appendix A at the end of the chapter for the corresponding SAS program. The primary result of Model 4 is the same as Model 3: differences in religiosity between MZ twins do not significantly predict delinquency, inconsistent with a causal hypothesis ($B_W = -0.008$). Second, the

within-pair effect is not significantly larger in DZ pairs than in MZ pairs ($B_{04} = 0.027$), indicating that the association between religiosity and delinquency is attributable to *environmental* factors that vary between families. Third, the residual variance shared by MZ twins (0.314) is larger than the variance shared by DZ twins (0.225), indicating that genetic factors account for residual variation in delinquency not accounted for by religiosity.

In the classical twin study, the three pieces of available information – between-and within-pair variances and zygosity – are reparameterized to yield the three components of the classical twin model: additive genetic influences (A), which are assumed to be perfectly correlated in MZ twin pairs and correlated at 0.5 in DZ twin pairs; shared environmental influences (C), which are environmental influences that make twins raised in the same home more similar, regardless of zygosity; and nonshared environmental influences (E), which are environmental influences that are unique to each twin and thus contribute to within-pair differences, plus measurement error. Differentiating these sources of variation depends on the relative similarity of MZ and DZ twins for a given phenotype. MZ twins are assumed to share all of their genes *and*, by definition, their shared environment; consequently, to the extent that MZ twins are not perfectly correlated for a phenotype, this is reflected in the estimate for E. To the extent that MZ twins, who share all of their genes, are more similar than DZ twins, who are assumed to share 50% of their genes, this will be reflected in the estimate of A. Finally, to the extent that the similarity of DZ twins exceeds half that observed in MZ twins, this will be reflected in the estimate of C.

Although most commonly discussed in terms of genetic and shared environmental contributions to an individual phenotype, we can also apply these same concepts to the *association* between two phenotypes, just as we could with the between and within variances in an analysis of MZ twins. To the extent the relation between religiosity and delinquency is truly causal, one would expect the same causal forces to be operating within families as between them. Our MZ twin analysis, however, indicated that the association between delinquency and religiosity was driven by between-family confounds rather than a “true” quasi-causal effect of religious involvement: Twin pairs who were more religious were less delinquent, but twins who differed in their religiosity did not differ in their delinquency. This model could be extended to ask whether the between-family confounds that drive the religiosity-delinquency association are genetic versus shared environmental in origin. It should be noted that both of these extensions are somewhat tangential to the central point of the analysis: testing whether the

quasi-causal hypothesis can be excluded. Nevertheless, this extension may be valuable in more fully characterizing the relation between predictor and outcome.

Assuming that the between-pair variance in MZ twins equals all genetic variance and all shared environmental variance, whereas the twin-pair variance in DZ twins equals one-half the genetic variance plus all shared environmental variance, simple arithmetic yields a more precise estimate for the genetic variance ($A = 2 \times (\text{MZ Twin Pair Var} - \text{DZ Twin Pair Var}) = 0.178$) and shared environmental variance ($C = \text{MZ Twin Pair Var} - A = 0.136$). The within-pair variance in MZ pairs is a direct estimate of the E variance (0.317).

In summary, random effects models of the relation between an environment and putative psychological outcome in twin pairs can be used to assess the following: (1) the extent to which within-twin pair differences in environmental experience predict differences in the outcome of interest, as would be expected under a causal hypothesis; (2) the extent to which the within-pair regression and the between-cluster regression are different, suggesting the operation of confounding variables; (3) the extent to which the within-pair regression in DZ pairs differs from MZ pairs, which suggests that the relevant confounds are genetic in origin; and (4) the extent to which the residual within-and between-pair variance components differ between MZ and DZ pairs, which suggests that genetic factors account for variation in outcome independent of the predictor of interest. The proposed mixed effects model, therefore, provides a rich characterization of how X and Y are related, with very minimal programming (six lines of SAS code).

Nevertheless, translating parameters that are specified in terms of pair means and deviations or between-and within-cluster variation into the components familiar in behavior geneticists – namely, additive genetics, shared environment, and nonshared environment – requires either post hoc arithmetic of the kind we have used here or rather elaborate weighting schemes (McArdle & Prescott, 2005). Moreover, whereas the current example clearly indicates that the relevant confounds were shared environmental in origin, there will obviously be cases in which the confounding variables are both genetic and environmental, and the proposed random effects model does not explicitly quantify genetic versus shared environmental selection effects. We now turn our attention to an analytic approach more commonly used in behavior genetics that addresses some of these shortcomings: structural equation modeling.

Structural Equation Models

Unstandardized ACE Regression

Since the development of powerful and relatively user-friendly structural equation modeling software, such as *Mplus* or *Mx*, twin data are most commonly analyzed using structural equation models, which, above and beyond offering a graphical means of representing the underlying equations, make it simpler to execute the reparameterizations of variances that we conducted in the preceding section using post hoc arithmetic. [Figure 8.1a](#) is a SEM representation of a twin regression, with delinquency predicted from religiosity in twin pairs. As was the case in the random effects regressions employed in the previous section, the dependence of the observations taken from the same twin pair is modeled explicitly, here simply by including estimates of the twin correlations for religiosity and delinquency, separately for MZ and DZ twins. Also similar to the random effect regressions, those MZ and DZ covariances can be reparameterized as ACE variances, as shown in [Figure 8.1b](#). Notably, the unstandardized regression of delinquency on religiosity, b_p , is the *same* in both figures; partitioning the variance of religiosity and delinquency has no consequences for the regression coefficient between them.

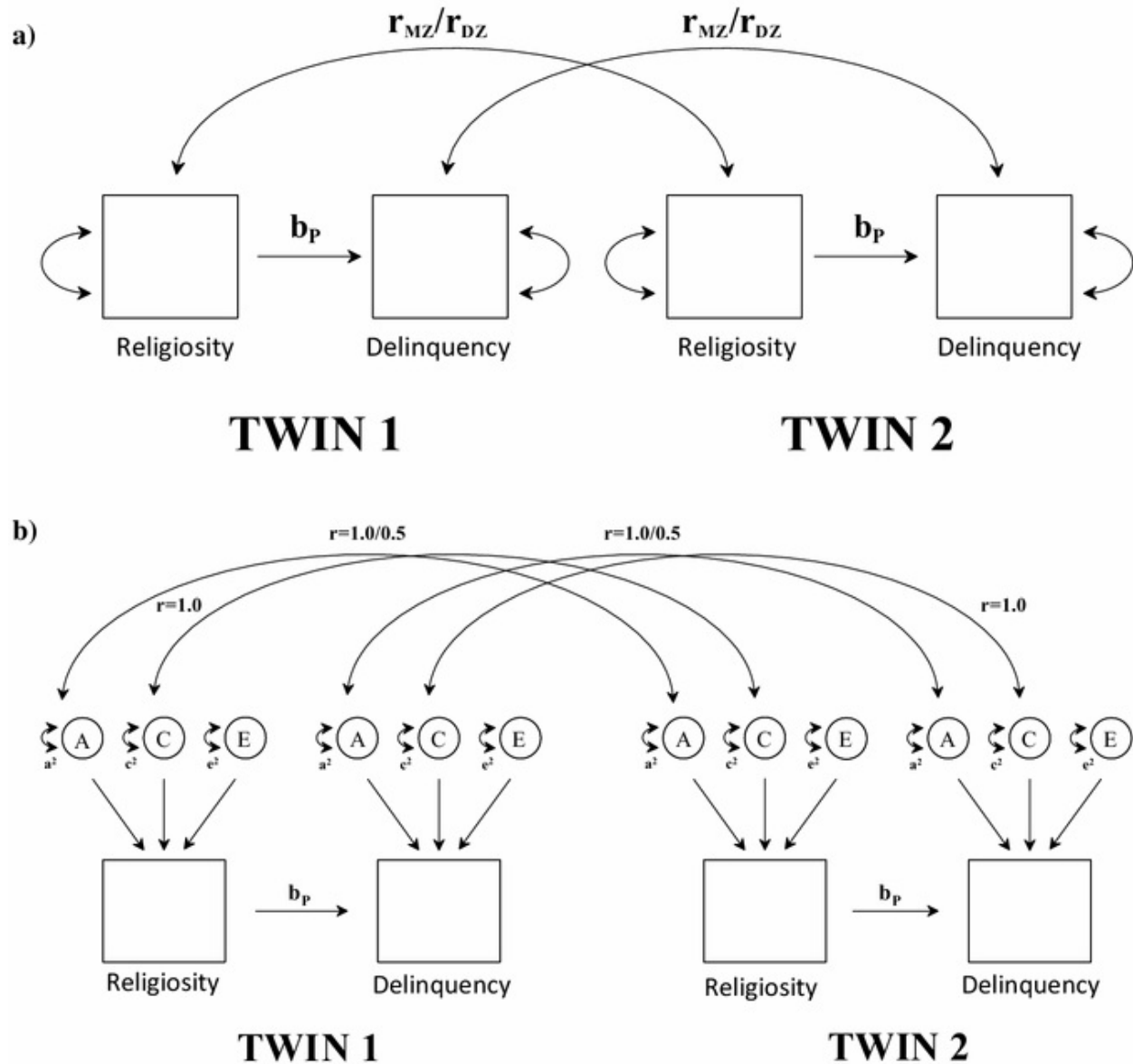


Figure 8.1. Phenotypic regression model in a pair of twins. Delinquency is regressed on religiosity, with equal regression coefficients b_P for each member of the pair. Curved paths represent twin correlations for delinquency and religiosity.

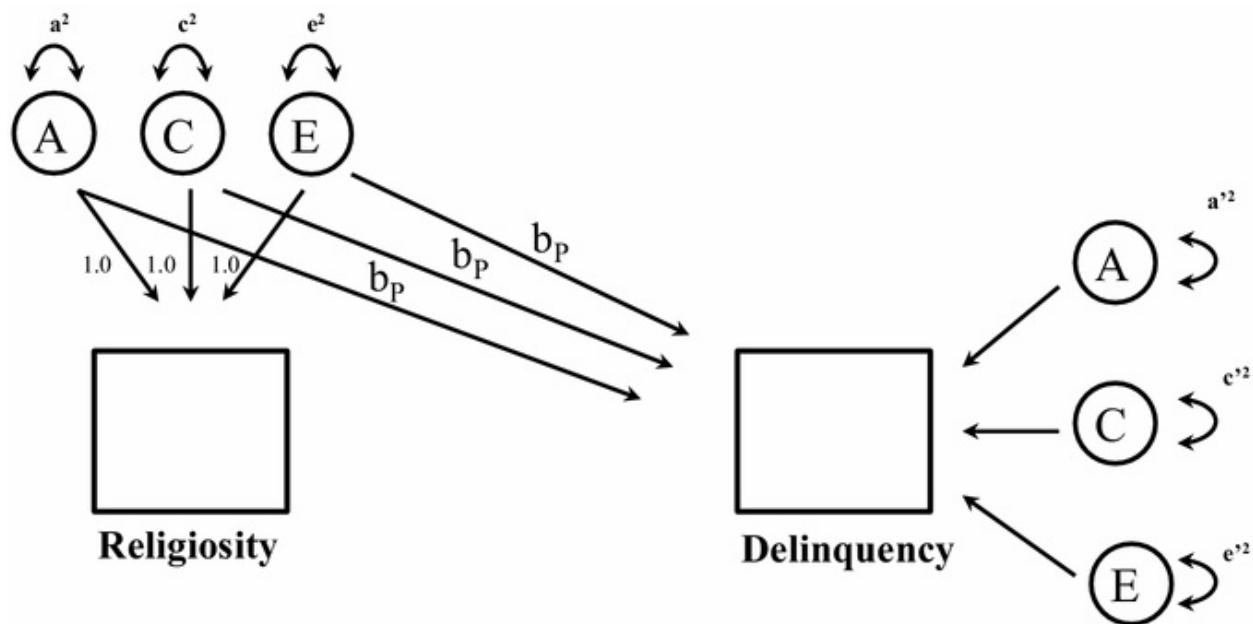


Figure 8.2. Phenotypic regression model in which twin covariances in MZ and DZ twins have been decomposed into biometric components; doing so has no effect on the regression b_P .

Nevertheless, in bivariate twin data of the kind required to fit the model in [Figure 8.1b](#), one can fit a model, illustrated in [Figure 8.2](#), in which separate regressions are estimated for Y on the three biometric components of X (only one twin is shown). If the true model is represented by [Figure 8.1a](#) (i.e., X causes Y), it is easy to see what will happen in a bivariate model. Because the phenotype X is the sum of its three unstandardized components A , C , and E , we will have

$$b_P(A + C + E) = b_P A + b_P C + b_P E \quad (8.5)$$

This shows that when the relationship between X and Y is causal at the level of the phenotype, the three unstandardized coefficients in a bivariate ACE regression model will be equal to each other and to the phenotypic causal parameter that underlies them.

The reasons for this equivalence are simple but easily misunderstood, and very important. Intuitions about the parameters of the classical twin model in the context of regression may be better served by temporarily suspending the classical “genes and environment” interpretation of the twin model in which the A term is identified with a latent construct called “the additive effect of genes,”

the C term is identified as “the shared environment” and the E term as the “nonshared environment” in favor of something more literal, as follows. The classical twin model can use covariances in identical and fraternal twins to partition a phenotype into three components, defined by their covariances between members of twin pairs. One (E) is uncorrelated between members of a pair, one (C) is perfectly correlated between members of a pair, and one (A) is correlated 1.0 in identical twins and 0.5 in fraternal twins. All three are components of the phenotype being analyzed, and their sum, $A+C+E$, is equal to the phenotype. Understood this way, the invariance of structural regression components in an unstandardized bivariate model of a causal relation is completely unsurprising. If one unit change in the *phenotype* of religiosity causes b_p units of change in delinquency, that will remain the case no matter what component of the phenotype one is examining, and whether one thinks of the origins of the component as genetic or environmental.

The foregoing discussion has not taken into account the problem at the heart of this chapter, namely that individuals are not randomly assigned to religiosity conditions, with the result that religiosity is potentially correlated with a host of unmeasured confounds that will then bias the phenotypic regression of delinquency on religiosity and vitiate the causal interpretation of simple regressions. In a manner entirely analogous to the random effects models developed earlier, with some key assumptions, twin models can provide insight into the nature of the confounding and any causal relations that remain. The key to the analysis is that any unmeasured confounds of religiosity can themselves be partitioned into ACE components, in the same way that confounds in a random effects model can be decomposed into a component correlated with variation in twin pair means and another component correlated with variation within pairs. The ACE components of the unmeasured confounds will then differentially bias the ACE regression coefficients with which they are associated. The model including the unmeasured confounds is illustrated in [Figure 8.3](#)

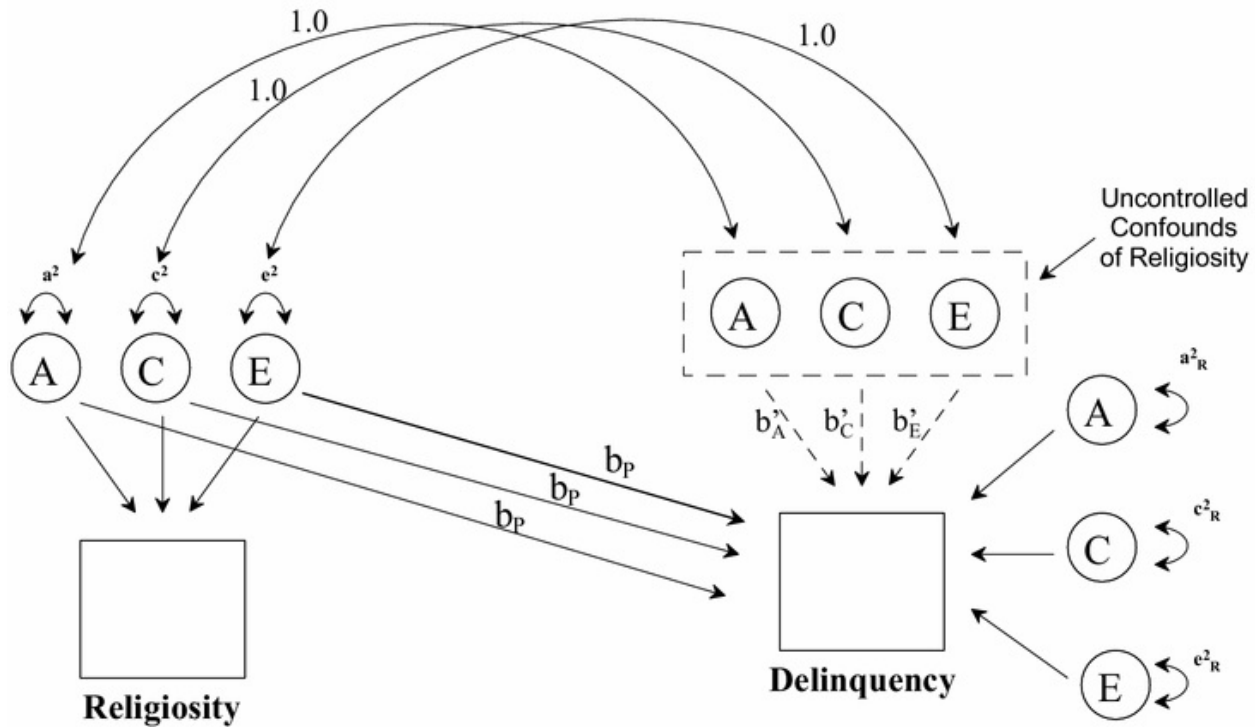


Figure 8.3. Genetically informed phenotypic regression, including unmeasured confounds (only one twin shown). Latent regressions along A, C, and E paths are equal to b_P , but observed regressions are biased by unmeasured A and C confounds in dotted box.

Figure 8.3 is identical to Figure 8.2, except for the addition of unobserved confounds inside the dotted rectangle, and dotted arrows with coefficients b'_A , b'_B and b'_E , which represent the causal effects of the confounds on the outcome. The confounds are perfectly correlated with the predictor of interest, because by definition the predictor and the uncontrolled confounds cannot be distinguished. In actual data, the ACE regression model is fit without reference to the unmeasured confounds, which then bias the unstandardized ACE regressions away from the true causal value b_P . In the model fit to the observed data,

$$b_A = b_P + b'_A; b_C = b_P + b'_C; b_E = b_P + b'_E \quad (8.6)$$

This situation would appear to present something of a dilemma, because there are three equations (one for each of the observed ACE regressions) and four unknowns (the true causal parameter and three ACE confounds). This is no small shortcoming, and should serve to remind us of the impossibility of

reaching strong causal inferences from nonexperimental data. Nonetheless, the ACE partitioning of the unmeasured confounds offers a partial solution. The A and C terms, representing variation in the confounds shared by identical twins, includes most of the plausible confounds of developmental causal relations. Socioeconomic status, parental education, culture, and, of course, genotype are all shared by identical twins, and therefore are not potential explanations of differences between members of a pair.

In contrast, the E component represents aspects of religiosity and its confounds that are uncorrelated within pairs of identical twins. It is more difficult – not impossible, just more difficult – to posit confounds of the relationship between religiosity and delinquency that are not shared by identical twins. At the very least, we can say that the E regression is free of the many genetic and environmental confounds that identical twins share. The partial solution to the underdetermination of Equation 8.6 (or, alternatively, the twin-based quasi-experimental solution to the unavailability of random assignment in studies of natural variation in humans) is as follows: To the extent we are willing to make the quasi-experimental assumption that relations within identical twin pairs are unconfounded by other variables, which is to say that b'_E is zero, then b_E estimates b_P , the true causal relation, while b_A and b_C estimate the sum of the causal coefficient plus the effects of confounds that vary in each of the domains, respectively.

The E regression estimates the causal effect of X on Y to the extent we can assume that differences within MZ pairs are not confounded by uncontrolled factors. We have already seen why this is a reasonable assumption to make. MZ twins are, to a first approximation, genetically identical, so differences in delinquency associated with differences in religiosity within pairs cannot be attributed to genetic differences between the pairs. When twins are reared together, they are roughly identical for a host of socioeconomic variables that might otherwise explain differences in delinquency: socioeconomic status, place of residence, and so on. It is, however, quite possible to think of uncontrolled third variables that might confound the within-pair association. For example, parents might choose to send one child to a religious school and the other to a public school. The child in the religious school would become more religious, and might also make friends that predisposed him or her lower levels of delinquency than the twin in the public school, even in the absence of any direct causal effect of religiosity on delinquency. Accepting the E regression as an estimate of the causal effect is therefore an assumption, or one might better say

an approximation, of the true state of affairs. In the absence of random assignment, strict inference of causality is simply impossible, and all twins offer is a quasi-experimental method for approaching it. As before, we prefer the term *quasi-causal* to describe a regression of Y on X (in either random effects or structural equation models) that has survived exposure to a genetically informed design that controls for genetic and shared environmental confounds.

The latent E component of religiosity is analogous to the within-pair deviation in religiosity ($x_{ij} - x_j$), entered as a covariate in mixed effects Models 2 – 4, above, and they offer the same inferential benefits and limitations. To understand this connection, consider again the case of MZ twins. They are necessarily identical for additive genetic and shared environmental factors; any difference between MZ twins is reflected in their scores on the latent E variable. That is, a higher score on the latent E variable indicates that an individual twin has higher levels of religiosity, relative to his or her co-twin. Consequently, the b_E path is directly analogous to the within-pair regression coefficient β_W . The b_E path asks, if Twin A has higher levels of religiosity than his or her co-twin (i.e., higher latent E scores), does Twin A also have lower levels of delinquency? Therefore, the regression on E is the key parameter of interest for testing a causal hypothesis about the relation between religiosity and delinquency.

Some readers may find this interpretation of the regression on E to be counterintuitive. Historically, there have been two important obstacles to a proper understanding of quasi-causal relations in behavior genetics, and the crucial role that is played by “nonshared environmental” differences within pairs of MZ twins. First, the three variance components of the classical twin model – additive genetics, shared and unshared environment – have been reified as “genes,” “shared environment,” and “nonshared environment” for so long that it is easy to forget that ultimately they are all just components of the variable being analyzed, specifically the predictor in a bivariate regression model, which in this case is phenotypic religiosity. That is to say, although according to a fairly restrictive set of assumptions the A component in a classical twin model of religiosity can be thought of in an abstract way as a standin for genetic variance in religiosity, the component itself *is* religiosity, *phenotypic* religiosity. The same is true of the C and E terms, which refer not so much to latent shared and nonshared environments underlying religiosity as to a portion of religiosity itself. By naming the latent variance that is not correlated between twins the nonshared environment, behavior geneticists have promoted as environmental what is really just within-pair variation in the phenotype of interest. Although

the *origin* of within-twin pair variation is, by definition, environmental influences (and measurement error) that make twins raised in the same family different, the *effect* of within-twin pair variation consists of the effect of *X* itself, confounded by the effects of any uncontrolled variables that also vary within pairs.

When describing the relation of a latent *E* variable with other variables in the model, many authors have mistakenly concluded that the *E* regression estimates only the role of nonshared environmental confounds, and have neglected that this relation is actually the best estimate of the phenotypic causal effects of *X* itself. Given that detection of this causal effect is usually the primary goal of the study, misinterpreting the term “nonshared” can sometimes snatch defeat from the jaws of victory in an otherwise successful study. For example, Pike, McGuire, Hetherington, Reiss, and Plomin (1996), although implying that their goal was to evaluate a causal hypothesis (“differential treatment affects adolescent adjustment”, p. 599), described their results as follows: “mothers’ negativity is significantly associated with depressive symptoms through nonshared environmental processes. (p. 597)” A reader could easily interpret this statement as meaning that some *other* environmental process, unique to each sibling, was responsible for both maternal negativity and depressive symptoms, rather than as evidence for a quasi-causal effect of mothers’ negativity on depression. Similarly, Spotts, Pederson, Neiderheiser, Reiss, Lichtenstein, Hansson, & Cederblad (2005) described a previous result (Reiss, Neiderheiser, Hetherington, & Plomin, 2000) as follows: “For example, more than half the correlation between mother’s positivity and child’s social responsibility is accounted for by genetic influences, with the remainder being accounted for by shared and nonshared environmental influences. (p. 339)” This statement makes it seem as though the association has been carved up into various types of confounds – a little attributable to genes, a little to social class or other between-family environmental differences, a little to peer groups or other within-family environmental differences – with nothing left over to comprise a causal relation. In fact, that “nonshared environmental influences” account for the “correlation between mother’s positivity and child’s social responsibility” means that within-pair differences in maternal positivity predicted within-pair differences in social responsibility – a quasi-causal relation. A third example is a report by McGue, Iacono, and Krueger (2006), whose stated goal was to evaluate whether early adolescent problem behavior is related to adult disinhibitory psychopathology via a causal mechanism, but who devoted a single sentence to describing, without comment, whether the twin with more problem behavior grew up to

have more psychopathology than his or her co-twin – “nonshared environmental factors accounted for the remaining 11% and 5% of the correlation [in females and males, respectively]” (p. 599).

Modeling Sequence

We previously described four ways in which the relation between an environment and putative psychological outcome in twin pairs can be characterized, each of which is explicitly characterized in the unstandardized bivariate ACE regression model. First, the extent to which within-twin pair differences in environmental experience predict differences in the outcome of interest, as would be predicted by a causal hypothesis, will be reflected in the b_e path. If b_e can be fixed to zero without significant loss of model fit, this provides disconfirmatory evidence regarding the quasi-causal hypothesis. Second, to the extent that the association between X and Y is confounded by variables that differ between unrelated individuals, the b_e path will differ from the b_a and b_c paths. If the three paths cannot be fixed to equality, this provides evidence that the quasi-causal association is confounded. Third, the relative magnitudes of b_a and b_c indicate the extent to which the confounding variables are genetic or environmental in origin. Finally, the extent to which genetic, shared environmental, and nonshared environmental factors contribute to residual variation in delinquency is directly estimated by the A, C, and E components of delinquency.

Nested SEMs can be compared using two measures of goodness-of-fit, Bayesian Information Criterion (BIC), and Root Mean Square Error of Approximation (RMSEA), as well as differences in χ^2 . BIC is an information-theoretic fit criterion that estimates the Bayes factor, the ratio of posterior to prior odds in comparisons of a model with a saturated one (Raftery & Richardson, 1996; Schwarz, 1978). BIC outperforms other fit criteria in its ability to discriminate between multivariate behavior genetic models, particularly for complex model comparisons in large samples, and is more robust to distributional misspecifications (Markon & Krueger, 2004). Interpretation of BIC values is entirely comparative, with lower values of BIC indicating better model fit. RMSEA measures error in approximating data from the model-per-model parameter (Steiger, 1990). RMSEA values of less than 0.05 indicate a close fit, and values up to 0.08 represent reasonable errors of approximation. Browne and Cudeck (1993) have argued that the RMSEA provides very useful

information about the degree to which a given model approximates population values. Differences in model χ^2 are themselves distributed as χ^2 , with df equal to the difference between the models' df.

We fit a series of five nested models to data from MZ and DZ twins in the program Mplus (Muthén & Muthén, 1998–2010). Results from the full model (Model 5) are summarized in Table 8.2. Of the total variance in religiosity, additive genetic effects accounted for approximately 24% [$4.30/(4.30+9.06+4.64)$], shared environmental influences for approximately 50%, and nonshared environmental influences for approximately 26%. Similarly, of the unique variance in delinquency, additive genetic effects accounted for approximately 25%, shared environment for 20%, and nonshared environment for 55%. Notice that the estimate for nonshared environmental variance ($E_{\text{Del}} = 0.334$) is equal (to the second decimal place) to the within-pair random effect, θ_{MZ} , estimated in the mixed effects Models 2–4. The regression onto E is not significantly different from zero ($b_e = 0.000$, 95% CI = -0.031 , 0.30), a result that falsifies the hypothesis that religiosity causes decreases in adolescent delinquency. The regression onto A is also not significantly different from zero ($b_a = 0.043$, 95% CI = -0.056 , 0.142), suggesting that genetic pathways are not a significant confound of the quasi-causal relation between religiosity and delinquency. The regression onto C is significantly different from zero ($b_c = 0.061$, 95% CI = 0.020 , 0.102), suggesting that environmental circumstances related to familial religiosity, such as parental education or socioeconomic status, account for the association between religiosity and delinquency.

Table 8.2. Results from Structural Equation Models of Religiosity and Delinquency

Parameters	Model 5: Full Model		Model 6: Regressions Equal		Model 7: No A or E Regression**	
	Unstandardized	Standardized	Unstandardized	Standardized	Unstandardized	Standardized
Variance Components of Religiosity						
Additive	4.30 (1.21)*	$h^2 = 24\%$	4.19 (1.21)*	$h^2 = 23\%$	4.30 (1.21)*	$h^2 = 24\%$
Genetic						
Shared	9.06 (.121)*	$c^2 = 50\%$	9.15 (1.21)*	$c^2 = 51\%$	9.04 (1.21)*	$c^2 = 50\%$
Environmental						
Non-Shared	4.64 (.39)*	$e^2 = 26\%$	4.66 (.40)*	$e^2 = 26\%$	4.64 (.39)*	$e^2 = 26\%$
Environmental						
Regression Paths						
A _{religion} → Delinquency	.043 (.050)	$\beta = .111$.036 (.005)*	$\beta = .091$	[0]	[0]
C _{religion} → Delinquency	.061 (.021)*	$\beta = .228$.036 (.005)*	$\beta = .135$.078 (.013)*	$\beta = .292$
E _{religion} → Delinquency	.000 (.016)	$\beta = -.001$.036 (.005)*	$\beta = .096$	[0]	[0]
Residual Variance Components of Delinquency						
Additive	.151 (.073)*	$h^2 = 25\%$.144 (.073)*	$h^2 = 23\%$.158 (.072)*	$h^2 = 27\%$
Genetic						
Shared	.124 (.060)*	$c^2 = 20\%$.136 (.058)*	$c^2 = 22\%$.103 (.058)	$c^2 = 17\%$
Environmental						
Non-Shared	.334 (.026)*	$e^2 = 55\%$.341 (.027)*	$e^2 = 55\%$.334 (.026)*	$e^2 = 56\%$
Environmental						
Model Fit Indices						
χ^2 (df, P)	30.36 (17, .02)		42.09 (19, .002)		31.64 (19, .03)	
$\Delta\chi^2$ (Δ df, P)	—		11.73 (2, .003)		1.28 (2, .527)	
CFI / TLI	.979 / .985		.963 / .977		.980 / .987	
RMSEA	.046		.057		.044	

* Significant at $P < .05$.

** Model accepted as the best representation of the data.

Model 6 tested whether environmental and genetic differences between families confound the association between religiosity and delinquency by fixing the regression paths to equality. This model resulted in a significant increase in χ^2 ($\Delta\chi^2 = 11.73$, $\Delta df = 2$, $P = 0.003$) and an increase in RMSEA, suggesting that the association is indeed confounded by between-family differences. The parameter estimates from Model 5, in which the only significant path between religiosity and delinquency was the regression on C, suggested that the relevant between-family confounds were shared environmental in origin. Model 7, then, fixed the regressions of delinquency on the A and E components to zero, allowing the association between religiosity and delinquency to be accounted for

entirely by environmental differences between families. Compared to the full model (Model 5), Model 7 did not fit significantly worse ($\Delta\chi^2 = 1.28$, $\Delta df = 2$, $P = 0.53$), indicating that shared environmental confounds could account for the association between delinquency and religiosity.

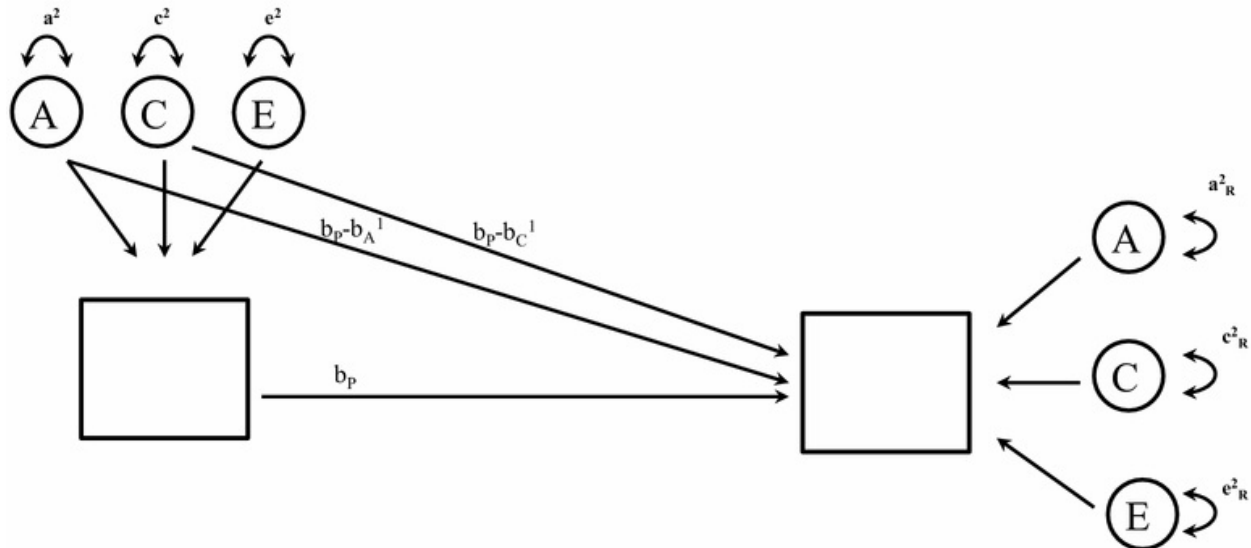


Figure 8.4. Genetically informed phenotypic regression, including unmeasured confounds (only one twin shown). Delinquency is regressed on religiosity along A, C and E pathways. E pathway estimates b_P .

The same conclusion is apparent, regardless of whether we use mixed effects models or structural equation models – religiosity is not causally related to delinquency, but families whose environmental circumstances make them the “type” of family to be religious have less delinquent children. This example, in particular, demonstrates the usefulness of ostensibly “genetic” designs in evaluating hypotheses about *phenotypic* causation in the presence of genetic and environmental confounds. The structural equation models have the added benefit of explicitly parameterizing genetic and shared environmental variance components, as well as separate genetic and shared environmental regressions. However, fitting these structural equation models requires substantially more programming than the mixed effects models (see Appendix B at the end of the chapter for program).

Alternative Parameterizations

The genetically informed regression model we have described provides a straightforward test of hypotheses regarding environmental causation, yet a scan

of the behavior genetics literature reveals a plethora of alternative parameterizations for bivariate twin or sibling data. While these parameterizations are, in most cases, mathematically indistinguishable (Loehlin, 1996), there are important conceptual differences among them that have implications for interpretative clarity. We include discussion of two reparameterizations that were not included in Loehlin (1996). One of these we recommend, the other not.

Genetically Informed Phenotypic Regression

A reparameterization of the bivariate regression model that we find to be interpretively useful is illustrated in Figure 8.4. As before, the outcome variable is regressed on the unstandardized A and C components of the predictor, but instead of also being regressed on the E term, it is regressed on the full phenotype of the predictor. This parameterization most closely approximates the substantive goal of research of this kind, which is to examine the phenotypic regression of Y on X while controlling for the between-pair genetic and environmental confounds contained in A and C, respectively.

Table 8.3. Results from Alternative Parameterization of Bivariate Cholesky (with Phenotypic Path from Religiosity to Delinquency)

	Model 8: Alternative Full Model (Phenotypic Path)	
Parameters	Unstandardized	Standardized
<i>Variance Components of Religiosity</i>		
Additive Genetic	4.30 (1.21)*	$h^2 = 24\%$
Shared Environmental	9.06 (.121)*	$c^2 = 50\%$
NonShared Environmental	4.64 (.39)*	$e^2 = 26\%$
<i>Regression Paths</i>		
A → Delinquency	0.42 (0.50)	$\beta = .112$

$A_{\text{religion}} \rightarrow \text{Delinquency}$.045 (.050)	$p = .112$
$C_{\text{religion}} \rightarrow \text{Delinquency}$.061 (.021)*	$\beta = .229$
Religion \rightarrow Delinquency	.000 (.016)	$\beta = -.002$

Residual Variance Components of Delinquency

Additive Genetic	.151 (.073)*	$h^2 = 25\%$
Shared Environmental	.124 (.060)*	$c^2 = 20\%$
NonShared Environmental	.334 (.026)*	$e^2 = 55\%$

Model Fit Indices

$\chi^2 (df, P)$	30.36 (17, .02)
$\Delta\chi^2 (\Delta df, P)$	—
CFI / TLI	.979 / .985
RMSEA	.046

* Significant at $P < .05$.

The interpretation of the coefficients in the genetically informed phenotypic regression model is to multiply through the coefficients for the ACE regression Equation 8.6. Redistributing, we obtain,

$$Y = b_{A'}A + b_{C''}C + (b_P b_{E'}) (A + C + E) \quad (8.7)$$

The sum of A , C , and E , of course, is simply X . So in this parameterization, the regression on the phenotype tests for the quasi-causal effect, on the assumption that there are no confounds within pairs of identical twins. The A and C regressions test the difference between the quasi-causal effect and the effects of

the A and C confounds, respectively. They are equal to zero when the quasi-causal relation is unconfounded. In this parameterization, the sequence of inferences is to test the phenotypic regression for the quasi-causal effect, then test whether b_A and b_C differ from 0 to test for the presence of confounding, then test the difference between b_A and b_C to determine if confounds are genetic or shared environmental in origin, and finally to test the residual variation in Y, as before.

The results of fitting the genetically informed phenotypic regression model are given in [Table 8.3](#). The phenotypic regression, analogous to the nonshared environmental regression in the previous models, is equal to zero, suggesting that once shared familial confounds have been controlled, there is no reason to hypothesize a quasi-causal relation between religiosity and delinquency. The values of the A and C regressions are the same as they were in the previous models, although this is not the usual result. In general, the A and C regressions in a phenotypic model will be equal to the *differences* between the A and C regressions and the E regression in an ACE regression model, which is to say they measure the magnitude of the bias introduced by the A and C confounds. In this example, however, the quasi-causal parameter is equal to exactly zero, so we conclude that the phenotypic regression is *all* confound, and the A and C regressions are the same in the ACE and phenotypic regression models.

Interpretation and Standardization

The reader experienced in fitting SEM models to twin data may have noticed that we have chosen to parameterize our models by estimating A, C, and E variances while fixing to 1.0 the paths from A, C, and E to the phenotype. We refer to this as an unstandardized parameterization. It is far more common to standardize the A, C, and E components of the predictor variable to a variance of 1, and estimate the paths to the observed variable, which we call a standardized parameterization. In univariate behavioral genetics, the choice between the standardized and unstandardized parameterizations is trivial. In the unstandardized parameterization, the variance accounted for by A, for example, is equal to the estimated A variance; in the standardized parameterization, it is equal to the square of the estimated path. The fit of the models is identical, and in a univariate design they are identical in interpretation.

In the context of multivariate regression models, however, these parameterizations have important conceptual differences. The complication

arises from the fact that the variances of the latent components of X have consequences for the regressions of Y on X . Return to Equation 8.7. Under conditions of no confounding, the unstandardized regression b_E estimates the quasi-causal regression b_P , and the three ACE regressions b_A , b_C , and b_E will be equal to each other. If we consider instead the standardized regressions β_A , β_C , and β_E , these relations no longer hold. We have, instead,

$$\beta_A = b_P \text{var}(A); \beta_C = b_P \text{var}(C); \beta_E = b_P \text{var}(E) \quad (8.8)$$

When the shared and nonshared variances are standardized, the regression of Y on the latent variables no longer depends solely on the structural coefficient relating the phenotypes, but is a function of this coefficient and the magnitude of the A , C , and E components of the predictor variable. Even when the phenotypic relation between X and Y is invariant – a structural, causally determined property – the magnitude of the latent variances can be expected to vary from population to population and study to study. In particular – as would be considered obvious in a typical regression context – the amount of variance in Y accounted for by the ACE components of X depends on the relative magnitudes of the ACE components.

Although the issue of standardization may at first seem to be a technical issue of biometric structural equation modeling, in fact it cuts to the heart of the behavior genetic enterprise. We will therefore make the point in a few different ways. One way to understand the difference between standardized and unstandardized regressions is as the difference between structural regression parameters (estimated by unstandardized regression coefficients) and variance explained (estimated by standardized regressions). Consider a concrete physical system that has been designed in a way that a change of one unit of X causes a change of two units of Y , with no other causes and no error. In any set of observations in which X varies, the unstandardized regression of Y on X will estimate the causal parameter – that is, two. (The standard error of the estimate will increase as the variance of X decreases, and of course no estimate can be computed when the variance of X becomes zero.) But what is the amount of variance in Y that is accounted for by X , or equivalently, the standardized regression coefficient of Y on X ? That quantity is equal to the square of the causal parameter multiplied by the variance of X . In conditions of high variability in X , it will account for a large amount of variability in Y , and in conditions of low variability in X it will account for a small amount of

variability; but the unstandardized parameter, two units of Y per unit of X , remains constant regardless of how variance in X may change.

The same kind of thinking applies to regressions with error of the kind we are considering here. Suppose that data on religiosity and delinquency were collected in the Netherlands rather than the United States, and that the phenotypic causal relationship between religiosity and delinquency was the same there as here, but for whatever reason in the Netherlands twins varied less in their religiosity. If the unstandardized model were fit, the investigator would gain insight into the similarities and differences between the American and Dutch populations: the quasi-causal relations represented by the unstandardized regressions (presumably the point of the study) would be identical, but the A , C , and E variance components would be generally be different from each other, and different in the Netherlands than they were in the United States. If the standardized model were fit, the structural invariance of the unstandardized parameter would be lost. The standardized regression parameters would be equal to the product of the invariant unstandardized parameters and the respective ACE variances. The investigator would observe that the regression coefficient linking adolescent rule-breaking to the E component of religiosity is larger in the Dutch study than in the American one, and, in the opaque language commonly encountered in multivariate behavior genetic research, conclude that the relation between religiosity and rule-breaking appears to be mediated along nonshared environmental pathways.

Another way of understanding the issue is in terms of metrics. When a phenotype is partitioned into unstandardized ACE components with estimated variances, all three components are expressed in the same units of X , so regression coefficients of Y on X are all expressed in the same units, that is, units of Y per units of X . If the latent components are standardized to have unit variance, however, their metric is no longer in units of X , but instead is equal to the standard deviations of each particular component. The A regression is in units of Y per standard deviation of A , the C regression in terms of the standard deviation of C , and so forth. In a biometric study, the magnitudes of these standard deviations are going to differ from each other, and across studies they will vary even more. Therefore the standardized regression coefficients relating Y to the biometric components of X will no longer be the same metric, and cannot be meaningfully compared.

Suppose a social psychologist interested in the fundamental attribution error set herself the task of determining the percentage of variance that the

fundamental attribution error (FAE; the FAE refers to a tendency to emphasize internal as opposed to situational explanations of the behavior of other people.) accounts for in some outcome. The research program would be a lost cause, because even assuming a fixed causal effect of the FAE that can be observed across situations, there is no invariant percentage of variance accounted for. In situations in which reliance on the FAE varies a great deal, it will account for a lot of variance in Y ; in situations where it hardly varies at all, it will account for very little. On the other hand, although the question of how much variance is accounted for by the fundamental attribution error is ill-conceived and not useful scientifically, it would be incorrect to conclude on this basis that the FAE itself was causally unimportant or that its effect could not be quantified. Based on either randomized experimentation or whatever quasi-experimentation can be cobbled together, causation is quantifiable by unstandardized regression coefficients, which are invariant against changes in the variance of X . The percentage of variance explained, quantified by standardized regression parameters, is not invariant, even when the underlying causal processes are.

To summarize: The structural constant underlying covariation between a cause and an effect is the unstandardized regression coefficient, expressing the units of Y caused by each unit of change in X . When regressions are standardized, they no longer estimate this quantity. Instead they estimate the variance in Y that is accounted for by X (if Y is also standardized, the variance explained will be a proportion), a quantity that depends on the structural parameter and the variance of X . Questions of how many units of reduction in delinquency are caused by an increase in one unit in religiosity, and whether that value is the same for religiosity itself as for the uncontrolled genetic and environmental traits that confound it, despite their myriad methodological and interpretive complexities, at least have an invariant correct answer that a diligent investigator can hope to estimate. It is true that relying on unstandardized coefficients forces us to take seriously the sometimes arbitrary units in which our constructs are measured, but that is ultimately a good thing, and in any event throwing the units away by standardizing makes regression analyses even harder to interpret. How much variance in delinquency is accounted for by religiosity and its ACE components has no invariant answer: It depends on the variability of delinquency and its biometric decomposition in a particular situation, and is not a worthwhile scientific question.

Heritability

We have saved until last the most important reason for the misapprehension of research methods in behavioral genetics: the concept of heritability, which we have intentionally delayed mentioning until this sentence. We have conducted a reasonably extensive review of behavioral genetic research methods without once referring to the construct that most defenders or critics of the field view as its central idea. (The most recent biologically oriented broadside against behavior genetics [Charney (2012)] refers to the object of its derision as “heritability studies.”) A full evaluation of heritability would take us far afield, but note that heritability is the proportion of phenotypic variability accounted for by the total effect of genotype (broad heritability) or by the additive effect of genotype (narrow heritability). Heritability is thus a standardized variance component, and as such it is not invariant as the genetic or phenotypic variance changes, so it is not a meaningful indicator of the causal effect of genotype on phenotype.

A good way to characterize the unstandardized multivariate models we have described in this chapter is that rather than being focused on the estimation of heritability, they are designed to estimate relationships between variables that are *invariant* as regards heritability. Religiosity and delinquency are heritable – everything is heritable or potentially so (Turkheimer, 2000) – and it would not be difficult to estimate heritability coefficients using the models we have described. The goal of our analyses, however, is not to estimate these coefficients, but rather to estimate the part of regression of delinquency on religiosity that is independent of their heritabilities. We want to estimate the extent to which delinquency is causally associated with religiosity above and beyond any variation in genetic background they may share. In the same way, the models control for any shared environmental variability that is common to religiosity and delinquency, again without attending to the magnitude of the shared environmental variance component. The only way to accomplish such invariance is to estimate unstandardized regressions, because the standardized alternative will depend in part on magnitudes of the biometric variances.

Molecular Genetic Approaches

Twenty years ago, most scientists studying the genetics of personality would not have predicted that the most crucial questions regarding causation would involve complex issues in the design and analysis of twin studies. As the Human Genome Project neared its completion, it was widely anticipated that the

availability of data from actual DNA, as opposed to the statistical inferences of quantitative genetics, would provide the causal, biologically based foundation that twin and family studies lacked. From the beginning, personality has played a signature role in the development of what are called “molecular” genetic methods for the study of behavior, providing some of the earliest successes but also some of the greatest frustrations. On the assumption that the molecular genetic methods are not yet as widely incorporated into the general body of research methods in personality, we review some of the basic techniques that are available. The review suggests that, in fact, the problems of scientific inference facing DNA-based studies of behavior turn out to have much in common with traditional family studies: The core scientific problem is still the inference of causation in a nonexperimental setting, and the contrasting of comparisons within and between family members continues to play a crucial role.

Linkage Analysis

The first DNA-based method that was applied to personality, linkage analysis operates within families. The word “linkage” refers the nonindependence of genetic loci that occur close to each other on a chromosome, a phenomenon called “linkage disequilibrium” (LD). In general, genes on different chromosomes are passed on independently, and crossover processes lead to independence for genes well separated on a single chromosome. Genes close together on the same chromosome will tend to be transmitted together, however. If within a family (either a complete pedigree or a pair of siblings and their parents), individuals who share a behavioral trait are also identical by descent (IBD) for a particular gene, it can be inferred that the trait in question is related to the gene or to another gene close to it on the chromosome.

Linkage analysis has been the earliest molecular method to be adopted in the study of behavior because it requires minimal knowledge of actual genetic sequence. The first linkage study of personality to be reported (Benjamin, Press, Maoz, & Belmaker, 1993) looked for linkage between the 16 PF and phenotypic color-blindness in 17 pairs of brothers of whom at least one was color-blind. Because color-blindness was known to be caused by a single X-linked locus, to the extent brothers concordant for color-blindness were more similar for personality, it would locate some of the variance in personality somewhere on the X chromosome. The results of the study foreshadowed much of what has happened since in the molecular genetics of personality. Of the 16 scales of the 16PF, one (Q2, self-sufficiency vs. group adherence) was more similar in the

brothers concordant for color-blindness compared to the discordant brothers: The authors conclude with a recommendation that the finding be replicated. The authors do not even mention statistical power, but with 17 sibling pairs it is obviously quite limited. More advanced linkage methods (e.g., Fullerton et al., 2003) use multiple markers to evaluate the probability of linkage continuously across the genome.

Linkage analysis has an important disadvantage as a means of studying personality. Although linkage can be detected with reasonable power using plausible sample sizes for single genes of large effect, it is almost certain that no such genes exist for normal variation in personality. For multiple genes of very small effect, which is just as certainly the situation that does obtain, the power to detect linkage is very small (Risch, 1990). At the same time, when genetic multiple markers are used in combination with multiple potential outcomes, the number of significance tests employed increases rapidly, meaning that both Type 1 and Type 2 error rates are often severely inflated. Another problem is that linkage analysis does not identify a single genetic locus that is associated with a phenotype, but rather a region of a chromosome within which variation in a gene is likely to be in LD with an outcome. Further analysis must be conducted (using association methods described later in the chapter) to establish the particular gene that is responsible for the linkage.

Candidate Gene Studies

In contrast to linkage analysis, which is conducted within families, most *candidate gene* or *association* studies are conducted between families. In its most basic form, the candidate gene study is about as simple as a study can be: A candidate genetic locus is mapped and a personality variable measured in a sample of unrelated individuals, and the outcome of the study is the correlation between the two. Most of the well-known molecular genetic studies of personality have been association studies. In what is still the most widely cited study of this kind, Ebstein *et al.* (1996) studied the relation between novelty seeking as measured by Tridimensional Personality Questionnaire (Cloninger, Svrakic, & Przybeck, 1993) and the D4 dopamine receptor gene (DRD4) in a sample 124 normal volunteers. Thirty-four participants who had at least one copy of the seven repeat allele scored about half a standard deviation higher than 90 who did not.

Whether or not the relation between DRD4 and novelty seeking has held up is a matter of meta-analytic controversy that we cannot review here (Kluger,

Siegfried, & Epstein, 2002; Munafo, Yalcin, Saffron, & Flint, 2008). What has become crystal clear, however, is that the effect size of half a standard deviation was unrealistically large. In the years since the heady early days of the genome project, it has become clear that associations between individual genetic loci and psychological outcomes are small and context-dependent, even when they appear to be statistically reliable. It has proven extremely difficult to screen reports of association studies for what is known as the “winner's curse” (Xiao & Boehnke, 2009): Faced with thousands of potential alleles, thousands of potential outcomes, intense pressure to publish, and a stringent peer-review system that prefers positive results, even a well-intentioned and honest community of scientists will produce effect sizes that are severely biased upward. The winner's curse is not exclusive to genomics; it is rampant in the behavioral sciences generally. It is just that the modern technology of genomics has brought with it an expectation of scientific rigor and replicability that the social sciences have long since gotten used to doing without (Turkheimer, 2011).

Genome-Wide Association Studies

The Gordian knot of methodology in association studies – myriad hypotheses, small effects, inadequate power, and results that seem to depend on context – has combined with the next wave of technological possibility in genomics to launch a new paradigm of DNA-based research. Genome-wide association studies, or GWAS, capitalize on the existence of single nucleotide polymorphisms (SNPs), individual segments of DNA that take only two values of the four (ATCG) that are possible. Upward of a million SNPs can be inexpensively assessed in a dense array across the genome. SNPs are not genes – they are indicators of genes, and associations between SNPs and outcomes are indicative of corresponding associations with genes at some location on the chromosome close enough to be in LD with the SNP.

With GWAS, it is possible to search for associations between personality and literally millions of locations in the genome. It quickly became apparent that for any trait of psychological interest, any one association would be tiny at best. The extremity of the inferential problems brought on by this situation has led to a radical restructuring of the way GWAS-based science is conducted. When candidate gene studies were first introduced, the theory-dependent process by which genes became candidates seemed like a tonic for the atheoretical gene searches of the early linkage era. But it turned out that the theory by which genes

were selected could not be separated from the chaotic problems of the winner's curse and publication bias. GWAS finally made these problems intractable, and the field has reversed course: rather than being guided by theory-driven hypotheses, GWAS is now conducted completely atheoretically, using highly stringent ($p < 10^{-8}$) significance levels to guard against Type I error. GWAS has produced important discoveries in the medical domain (Visscher & Montgomery, 2009), but it has been disappointing at best in the behavioral sciences, and particularly so for personality (Munafo, Clark, Moore, Payne, Walton, & Flint, 2003).

There is another threat to the validity of association studies, usually identified with GWAS but in fact relevant to all studies of genetic association: population stratification, sometimes called the “chopsticks gene” problem (Hamer, 2000). Here is how one could find a gene associated with the use of chopsticks. Assemble a sample consisting of half North American and half Japanese participants, and identify a gene – any gene will do – that occurs more frequently in the Japanese than in the North Americans. Assuming the Japanese are more likely to use chopsticks, eating style will necessarily be correlated with variation in the gene. Population stratification is usually controlled either by ensuring ethnic homogeneity in the sample or by using statistical methods like principal component analysis to identify ethnic dimensions in the SNPs and then controlling for them statistically. The problem, however, extends beyond the confines of ethnic identification (Turkheimer, 2011). In the context of GWAS, culture of origin is an instance of the core problem that we identified in the analysis of twin studies – a shared environmental confound. In the chopsticks example, the problem is not that the candidate gene and chopsticks may not actually be associated, because their correlation is a simple statistical fact. The problem is that the statistical association is not an indicator of an actual causal pathway from the gene to the chopstick, in exactly the same sense that holds for an association between religiosity and delinquency. In fact, in a mixed sample of Japanese and North American youth in which the Japanese are more religious and less delinquent, there would be a statistical association that would probably not be indicative of a causal relation.

What can be done to rescue causal inference under such extreme nonexperimental conditions? Resisting the temptation to say “nothing,” one interesting possibility is to return to within-and between-families approach that we endorsed as a structure for causal inference in twin studies, and which formed the basis for linkage studies in the early days of molecular genetics. An association between a gene and chopstick use within families (the sibling with

the gene used chopsticks; the one without the gene did not) effectively rules out population stratification as an alternative explanation. Designs combining within-and between-families genetic associations and the statistical methods to analyze them are available and predate GWAS (Fulker, Cherny, Sham, & Hewitt, 1999), but have not been widely employed for personality. One exception is a study by Middeldorp, de Geus, Beem, Lakenberg, Hottenga, Slagboom, and Boomsma (2007), which studied the relation between the serotonin transporter gene (5-HTTLPR) and neuroticism, anxiety, and depression in a sample of 1,804 twins, both sibling and parents. Only two of the eighteen within-family tests reached a significance level of $p < .05$, leading the authors to reject the hypothesis of a causal relation between 5-HTTLPR and the outcomes.

The molecular genetics of personality has reached a conundrum. One can design “theory-driven” studies within and between families, which control for a subset of potential confounds of genomic causation, but which are unavoidably contaminated by data exploration and the winner's curse, cherry picking of results, and publication bias. These studies wind up looking like non-genomic social science: locally interesting but frustratingly noncumulative. Or, one can opt for GWAS of massive populations with tiny p levels, atheoretical by design and blind to the possibility of noncausal confounds, hoping for a few reliably significant effects that collectively account for a few percent of the variance at best, and which have not, in the behavioral sciences at any rate, produced substantive causal science. What would seem to be the logical compromise – GWAS of enormous samples of siblings – simply isn't practical.

Genome-Wide Complex Trait Analysis

We close this section with a consideration of the newest molecular genetic method. Genome-wide complex trait analysis (GCTA) uses GWAS data in a novel way that closes the methodological gap between quantitative and molecular genetics (Visscher, Yang, & Goddard, 2010; Yang et al., 2010; see also Turkheimer, 2012). In a sample of individuals from whom SNP chips have been obtained, pairwise coefficients are computed that quantify the degree of genetic similarity between pairs of individuals across the SNPs. These coefficients, which are analogous to the coefficients of genetic relatedness in twin and family studies (e.g., siblings are on average 50% genetically related), are generally close to zero, and in fact pairs with coefficients higher than 2.5% are usually eliminated as genetically related. Once the genetic similarity matrix

is obtained, one can compute the relationship between the degree of SNP-based genetic similarity and the degree of similarity in the trait of interest, obtaining a proportion of variance that is essentially a heritability coefficient computed in a sample of unrelated individuals. These heritabilities are generally smaller than those computed from family members, but considerably larger than the percentage of variance that can be obtained by adding up the effects of individual SNPs or genes.

Vinkhuyzen *et al.* (2012) conducted GCTA of neuroticism and extraversion scores in a sample of approximately 12,000 individuals collected from several research centers. Across the centers, the traits had quantitative heritabilities (computed in the usual way) of .4 to .45. In contrast, 6% of the variation in neuroticism and 12% of the variation in extraversion could be explained by SNP-based similarity. These proportions are somewhat smaller than they have been for other traits, like height and intelligence, for which almost half of the phenotypic heritability has been recovered from the SNPs, although 6% and 12% are still significantly more than the 1% that can be recovered by quantifying the effects of individual genes. It should be emphasized that the methodology of GCTA is in fact much more similar to a twin study than it is to a GWAS. No distinction is made between the effects of individual SNPs, and no inference of the causal effects of individual SNPs is even attempted (Turkheimer, 2012).

Conclusions and Recommendations

Historically, it is undeniable that behavior genetics has progressed from Galtonian ideas about “nature and nurture,” by way of supportable notions of heritability in animal breeding, to a long era of concern, if not outright obsession, with the values of heritability coefficients for human individual differences. Other than the important task of disconfirming any remnants of blank-slate environmentalism mistakenly held over from previous eras of behaviorism or psychoanalysis, this effort was in our view not especially productive. Heritability is greater than zero for all individual differences, and takes a determinate value for none of them. Figuring out how “genetic” traits are, either in absolute terms or relative to each other, is a lost cause: Everything is genetic to some extent and nothing is completely so. There is little more to be said.

But despite the endless assertions of heritability and the similarly endless denunciations of behavior genetic studies and their conclusions, both of which

continue unabated to this day, we contend that heritability was never the most important motivation for human of behavioral genetics. Instead, behavioral genetics is justified by the simple observation that there is more than one reason why differences among people are correlated with each other, either within individual lives or across genetically related individuals. There are genetic as well as environmental reasons why extraverted mothers have extraverted children, or why religiously committed youth are less likely to become delinquent. Any question worth asking about the behavioral genetics of personality comes down to a question about *why* two or more traits are related to each other, and like any other kind of association-based psychology, such questions are ultimately about whether and how one trait causes another. Once that point is conceded, a huge segment of nonexperimental human psychology threatens to collapse unless genetically informative designs can be called on to support it, and such designs turn out not to depend on point estimates of heritability at all; indeed, their correct analysis relies on methods that are invariant with regard to changes in the genetic and environmental variability of individual differences.

Our analysis leads to a several specific recommendations for the conduct of genetically informed research in personality, and we will close by enumerating them.

1. Behavioral genetic investigations of relations among personality variables or between personality and exogenous variables should begin with an observation of a phenotypic association, which will usually be uncontrolled by random assignment. The goal of the genetically informed part of the analysis is to expose the causal basis of the phenotypic association to risk of disconfirmation.
2. Regression-based genetically informative analyses can be conducted more or less equivalently using multilevel models, structural equation models, or a combination of the two. Multilevel models usually have the advantage of being easier to code, whereas structural equation models have the advantage of greater flexibility, especially in their ability to re-parameterize random variances into the familiar biometric components.
3. Although behavior genetic designs are commonly thought of as a means of identifying and controlling genetic effects, shared environmental confounds are often equally important threats to causal hypotheses. If neuroticism is associated with poorer school performance, but living in

a violent neighborhood contributes to both, the shared environmental effect of neighborhood is an alternative, noncausal environmental explanation of the phenotypic association.

4. Causal hypotheses are almost always about phenotypic relations among variables, not relations among abstract variance components presumably representing genetic and environmental processes underlying observed behavior. Nonshared environmental regressions are usually the best available estimate of causal relations among uncontrolled variables, because neither genes nor shared environments can account for them. The nonshared environment plays a special role in genetically informed social scientific methodology because it encompasses associations among variables that cannot be accounted for by shared genes or environments, and are thus more plausible instances of phenotypic causation.
5. Notwithstanding the aforementioned, uncontrolled associations within identical twin pairs are not immune from confounds, and behavior genetic methodology is ultimately just another quasi-experimental tool in the social scientific workshop. Once phenotypic associations have survived exposure to analyses of genetic and shared environmental confounds, confidence in the causal relation may increase, but it is not proven. We prefer the term “quasi-causal” to describe the hypothesis that remains.
6. From the point of view of understanding the relationship between behavior genetics and the rest of naturalistic developmental psychology, the inferential imperfection of genetically informed designs is a good thing. Too often, proponents and detractors of behavior genetics describe the discipline as though it were somehow alien to the rest of social scientific methodology, generating either robustly scientific or falsely reductionist genetic counterhypotheses to psychological theories of human development. Neither is true. Behavior genetics is only a threat to psychological theories in the same sense that the cross-lagged panel design is. Yes, behavior genetic designs can sometimes make it harder to believe in causal hypotheses (Turkheimer & Waldron, 2000), but that is as it should be, and ultimately behavior genetics is no more probative of causal relations than any other quasi-experimental method.
7. To a surprising degree, issues of standardization are crucial to placing behavior genetic methodology on a strong foundation. Since Tukey, it has been well-understood that only unstandardized regression coefficients provide invariant estimates of causal relations in the face of

changes in the variances of predictor and outcome, but unfortunately those old insights have been lost in contemporary practice that still relies on correlations and standardized “beta weights.” Accounting for variance and explaining causation are two different and ultimately independent enterprises, and science is almost always properly concerned with the latter.

8. As Tukey famously said about correlation coefficients, we believe that the world would be a less confusing and contentious place without heritability coefficients, at least if one is concerned with a more complex and uncontrollable aspect of behavior than, say, milk production in cows. As with the heritability of milk production, the heritability of neuroticism informs us that we could selectively breed for human neuroticism if we wanted to, but fortunately we do not. Genes, in the very abstract sense in which the term is used in human quantitative genetics, influence neuroticism, and this will generally ensure that the heritability of neuroticism is not zero. Beyond that the numerical value of heritability is indeterminate, and the question of “how important” genes are to differences in neuroticism has no meaningful answer.
9. Skepticism about the utility of heritability coefficients should not be a basis for believing that genetic variance in neuroticism does not matter. It does matter, because familial variance in neuroticism and its familial covariance with other traits are alternative explanations of causal hypotheses about its phenotypic origins and consequences.
10. Molecular genetic methods have added to the tools available to behavior geneticists, but they have not replaced twin studies and quantitative genetic statistical methods. Just as we have contended that the goal of twin studies was never to quantify the magnitude of genetic effects on phenotypes, the goal of molecular genetics is not to discover the individual genes that underlie differences in personality. Instead, the goal of both quantitative and molecular genetics is to aid in the identification of causal processes in development, and in that regard molecular genetics faces many of the same problems as quantitative genetics, often in even more intractable forms.
11. Viewed in this way – as a quasi-experimental method that sits alongside many others – behavior genetic research methods can be seen for what they truly are rather than as the threatening or naïve stereotypes that are often represented. Behavior genetics is not a radical, reductive alternative to psychological explanation of behavior, as earlier

critics once feared (Lewontin, Rose, & Kamin, 1984), and it is not a poorly specified, dumbed-down version of the astounding understanding of genomics we have achieved at a biological level of analysis, as more recent critics have contended (Charney, 2012). Behavior genetics is not oversimplified genomics any more than nonexperimental developmental psychology is oversimplified developmental biology. Behavior genetics is ordinary social science, with all the problems that come with a lack of experimental control; it is, however, social science conducted a little more carefully, analyzed with a little more realism about why individual differences are associated with each other, and interpreted with a little more skepticism about the vagaries of correlation and causation.

References

- Alpert, H. (1939). Emile Durkheim and sociologismic psychology. *American Journal of Sociology*, 45(1), 64–70.
- Baier, C. J., & Wright, B. R. E. (2001). “If you love me, keep my commandments”: A meta-analysis of the effect of religion on crime. *Journal of Research in Crime and Delinquency*, 38(1), 3–21.
- Benjamin, J., Press, J., Maoz, B., Belmaker, R. H. (1993). Linkage of a normal personality trait to the color-blindness gene: Preliminary evidence. *Biological Psychiatry*, 34(8), 581–583.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage Publications.
- Burt, S. A. (2009). A mechanistic explanation of popularity: Genes, rule breaking, and evocative gene-environment correlations. *Journal of Personality and Social Psychology*, 96(4), 783–794.
- Carlin, J. B., Gurrin, L. C., Sterne, J. A., Morley, R., & Dwyer, T. (2005). Regression models for twin studies: A critical review. *International Journal of Epidemiology*, 34(5), 1089–1099.
- Cattell, R. B. (1957). *Personality and motivation structure and measurement*. Oxford: World Book Company.
- Charney, E. (2012). Behavior genetics and postgenomics. *Behavioral and Brain*

Sciences, 35(5), 331–358.

- Cloninger, C. R., Svrakic, D. M., & Przybeck, T. R. (1993). A psychobiological model of temperament and character. *Archives of General Psychiatry*, 50(12), 975–990.
- Ebstein, R. P., Novick, O., Umansky, R., Priel, B., Osher, Y., Blaine, D., Bennett, E. R., Nemanov, L., Katz, M., & Belmaker, R. H. (1996). Dopamine D4 receptor (D4DR) exon III polymorphism associated with the human personality trait of Novelty Seeking. *Nature Genetics*, 12, 78–80.
- Ellison, C. G., & Sherkat, D. E. (1995). The “semi-involuntary institution” revisited: Regional variations in church participation among Black Americans. *Social Forces*, 73(4), 1415–1437.
- Farrington, D. P. (2005). Childhood origins of antisocial behavior. *Clinical Psychology & Psychotherapy*, 12(3), 177–190.
- Federal Bureau of Investigation. (2004). *Crime in the United States: 2004*. Washington, DC: U.S. Government Printing Office.
- Fulker, D. W., Cherny, S. S., Sham, P. C. & Hewitt, J. K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics*, 64(1), 259–267.
- Fullerton, J., Cubin, M., Tiwan, H., Wang, C., Bomhra, A., Davidson, S., Miller, S., Fairburn, C., Goodwin, G., Neale, M. C., Fiddy, S., Mott, R., Allison, D. B., & Flint, J. (2003). Linkage analysis of extremely discordant and concordant sibling pairs identifies quantitative-trait loci that influence variation in the human personality trait neuroticism. *American Journal of Human Genetics*, 72(4), 879–890.
- Gosling, S. D., & John, O. P. (1999). Personality dimensions in nonhuman animals: A cross-species review. *Current Directions in Psychological Science*, 8(3), 69–75.
- Hamer, D. (2000). Beware the chopsticks gene. *Molecular Psychiatry*, 5(1), 11–13.
- Harden, K. P., Mendle, J., Hill, J. E., Turkheimer, E., & Emery, R. E. (2008). Rethinking timing of first sex and delinquency. *Journal of Youth and Adolescence*, 37(4), 373–385.

- Harris, K. M. (2011). *Design features of Add Health*. Chapel Hill, NC: Carolina Population Center at the University of North Carolina at Chapel Hill. Available at <http://www.cpc.unc.edu/projects/addhealth/data/guide>.
- Harris, K. M., Halpern, C. T., Smolen, A., & Haberstick, B. C. (2006). The National Longitudinal Study of Adolescent Health (Add Health) twin data. *Twin Research and Human Genetics*, 9, 988–997.
- Hopper, J. L., & Seeman, E. (1994). The bone density of female twins discordant for tobacco use. *New England Journal of Medicine*, 330(6), 387–392.
- Jackson, J. J., Thoemmes, F., Jonkmann, K., Lüdtke, O., & Trautwein, U. (2012). Military training and personality trait development does the military make the man, or does the man make the military? *Psychological Science*, 23(3), 270–277.
- Kazdin, A. E. (1997). Parent management training: Evidence, outcomes, and issues. *Journal of the American Academy of Child and Adolescent Psychiatry*, 36(10), 1349–1356.
- Koenig, L. B., McGue, M., Krueger, R. F., & Bouchard, T. J. (2005). Genetic and environmental influences on religiousness: Findings for retrospective and current religiousness ratings. *Journal of Personality*, 73(2), 471–488.
- Laub, J. H., & Sampson, R. J. (1993). Turning points in the life course: Why change matters to the study of crime. *Criminology*, 31(3), 301–325.
- Lewontin, R. C., Rose, S. R., & Kamin, L. J. (1984). *Not in our genes: Biology, ideology, and human nature*. New York: Pantheon Books.
- Littlefield, A. K., Sher, K. J., & Wood, P. K. (2009). Is “maturing out” of problematic alcohol involvement related to personality change? *Journal of Abnormal Psychology*, 118(2), 360.
- Loehlin, J. C. (1996). The Cholesky approach: A cautionary note. *Behavior Genetics*, 26(1), 65–69.
- Mann, V., De Stavola, B. L., & Leon, D. A. (2004). Separating within and between effects in family studies: An application to the study of blood pressure in children. *Statistics in Medicine*, 23(17), 2745–2756.
- Markon, K. E., & Krueger, R. F. (2004). An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic

- models. *Behavior Genetics*, 34(6), 593–610.
- McArdle, J. J., & Prescott, C. A. (2005). Mixed-effects variance components models for biometric family analyses. *Behavior Genetics*, 16(1), 163–200.
- McGue, M., Iacono, W. G., & Krueger, R. (2006). The association of early adolescent problem behavior and adult psychopathology: A multivariate behavioral genetic perspective. *Behavior Genetics*, 36(4), 591–602.
- Middeldorp, C. M., deGeus, E. J. C., Beem, A. L., Lakenberg, N., Hottenga, J., Slagboom, P. E., & Boomsma, D. I. (2007). Family based association analyses between the serotonin transporter gene polymorphism (5-HTTLPR) and neuroticism, anxiety and depression. *Behavior Genetics*, 37(2), 294–301.
- Miles, D. R., & Carey, G. (1997). Genetic and environmental architecture of human aggression. *Journal of Personality and Social Psychology*, 72(1), 207–217.
- Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review*, 100(4), 674–701.
- Morley, R., Moore, V. M., Dwyer, T., Owens, J. A., Umstad, M. P., & Carlin, J. B. (2005). Association between erythropoietin in cord blood of twins and size at birth: Does it relate to gestational factors or to factors during labor or delivery? *Pediatric Research*, 57, 680–684.
- Munafo, M. R., Clark, T. G., Moore, L. R., Payne, E., Walton, R., & Flint, J. (2003). Genetic polymorphisms and personality in healthy adults: A systematic review and meta-analysis. *Molecular Psychiatry*, 8(5), 471–484.
- Munafo, M. R., Yalcin, B., Saffron, A. W., & Flint, J. (2008). Association of the dopamine D4 receptor (DRD4) gene and approach-related personality traits: Meta-analysis and new data. *Biological Psychiatry*, 63(2), 197–206.
- Muthén, B., & Muthén, L. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.
- Neale, M. C., & Maes, H. H. (2007). *Methodology for genetic studies of twins and families*. Dordrecht, The Netherlands: Kluwer Academic.
- Oltmanns, T. F., & Turkheimer, E. (2009). Person perception and personality pathology. *Current Directions in Psychological Science*, 18(1), 32–36.

- Pearce, L. D., & Axinn, W. G. (1998). The impact of family religious life on the quality of mother-child relations. *American Sociological Review*, 63(6), 810–828.
- Penke, L., Denissen, J. J. A., & Miller, G. F. (2007). The evolutionary genetics of personality. *European Journal of Personality*, 21(5), 549–587.
- Pike, A., McGuire, S., Hetherington, E. M., Reiss, D., & Plomin, R. (1996). Family environment and adolescent depressive symptoms and antisocial behavior: A multivariate genetic analysis. *Developmental Psychology*, 32(4), 590–603.
- Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2013). *Behavioral genetics* (6th ed.). New York: Worth Publishers.
- Quinn, P. D., Stappenbeck, C. A., & Fromme, K. (2011). Collegiate heavy drinking prospectively predicts change in sensation seeking and impulsivity. *Journal of Abnormal Psychology*, 120(3), 543.
- Quinsey, V. L., Skilling, T. A., Lalumière, M. L., & Craig, W. M. (2004). *Juvenile delinquency: Understanding the origins of individual differences*. Washington, DC: American Psychological Association.
- Raftery, A. E., & Richardson, S. (1996). Model selection for generalized linear models via GLIB, with application to epidemiology. In D. A. Berry & D. K. Stangl (Eds.), *Bayesian biostatistics* (pp. 321–354). New York: Marcel Dekker.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.
- Reiss, D., Neiderhiser, J. M., Hetherington, E. M., & Plomin, R. (2000). *The relationship code: Deciphering genetic and social influences on adolescent development*. Cambridge, MA: Harvard University Press.
- Risch, N. (1990). Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *American Journal of Human Genetics*, 46(2), 229–241.
- Sampson, R. J., Morenoff, J. D., & Gannon-Rowley, T. (2002). Assessing “neighborhood effects”: Social processes and new directions in research. *Annual Review of Sociology*, 28, 443–478.

- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Scott, J. P., & Fuller, J. L. (1965). *Genetics and the social behavior of the dog*. Chicago: University of Chicago Press.
- Shipley, B. A., Weiss, A., Der, G., Taylor, M. D., & Deary, I. J. (2007). Neuroticism, extraversion, and mortality in the UK Health and Lifestyle Survey: A 21-year prospective cohort study. *Psychosomatic Medicine*, 69(9), 923–931.
- Smith, D. N. (2003). *Hinduism and modernity*. Malden, MA: Blackwell Publishing.
- Spitz, E., Moutier, R., Reed, T., Busnel, M. C., & Marchaland, C. (1996). Comparative diagnoses of twin zygosity by SSLP variant analysis, questionnaire, and dermatoglyphic analysis. *Behavior Genetics*, 26(1), 55–63.
- Spotts, E. L., Pederson, N. L., Neiderhiser, J. M., Reiss, D., Lichtenstein, P., Hansson, K., & Cederblad, M. (2005). Genetic effects on women's positive mental health: Do marital relationships and social support matter? *Journal of Family Psychology*, 19(3), 339–349.
- Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25(2), 173–180.
- Tellegen, A., Lykken, D. T., Bouchard, T. J., Wilcox, K. J., Segal, N. L., & Rich, S. (1988). Personality similarity in twins reared apart and together. *Journal of Personality and Social Psychology*, 54(6), 1031–1039.
- Tukey, J. W. (1954). Causation, regression, and path analysis. In O. Kempthorne, T. A. Bancroft, J. W. Gowen, & J. L. Lush (Eds.), *Statistics and mathematics in biology* (pp. 35–66). Ames: Iowa State University Press.
- Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, 24(2), 83–91.
- Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science*, 9(5), 160–164.
- Turkheimer, E. (2011). Genome wide association studies of behavior are social science. In K. S. Plaisance & T. A. C. Reydon (Eds.), *Philosophy of behavioral biology* (pp. 43–64). New York: Springer.

- Turkheimer, E. (2012). Still missing. *Research in Human Development*, 8(3–4), 227–241.
- Turkheimer, E., & Waldron, M. (2000). Nonshared environment: A theoretical, methodological, and quantitative review. *Psychological Bulletin*, 126(1), 78–108.
- Vinkhuyzen, A. A. E., Pedersen, N. L., Yang, J., Lee, S. H., Magnusson, P. K. E., *et al.* (2012). Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. *Translational Psychiatry*, 2, e102. doi:10.1038/tp.2012.27
- Visscher, P. M., & Montgomery, G. W. (2009). Genome-wide association studies and human disease: From trickle to flood. *Journal of the American Medical Association*, 302(18), 2028–2029.
- Visscher, P. M., Yang, J., & Goddard, M. E. (2010). A commentary on ‘common SNPs explain a large proportion of the heritability for human height.’ *Twin Research and Human Genetics*, 13, 517–524.
- Wallace, J. M., Brown, T. N., Bachman, J. G., & Laveist, T. A. (2003). The influence of race and religion on abstinence from alcohol, cigarettes and marijuana among adolescents. *Journal of Studies on Alcohol*, 64(6), 843–848.
- Wilcox, W. B. (1998). Conservative protestant childrearing: Authoritarian or authoritative? *American Sociological Review*, 63(6), 796–809.
- Wilson, J., & Sherkat, D. E. (1994). Returning to the fold. *Journal for the Scientific Study of Religion*, 33(2), 148–161.
- Xiao, R., & Boehnke, M. (2009). Quantifying and correcting for the winner's curse in genetic association studies. *Genetic Epidemiology*, 33(5), 453–462.
- Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K. *et al.* (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7), 565–569.

Appendix A: Sas Code

f11 =

Delinquency factor score at Wave I

religsum =

Individual religiosity sum score

religavg =

Twin pair average religiosity

religdev =

Individual deviation from twin pair average religiosity

MZ =

0 or 1; Monozygotic twin pair

Model 1

```
proc mixed data=mztwin12 method= ml covtest noclprint;  
class pair;  
model f11 = religsum solution;  
random intercept subject=pair type=vc;  
run;
```

Model 2

```
proc mixed data=mztwin12 method= ml covtest noclprint;  
class pair;  
model f11 = religsum religavg solution;  
random intercept subject=pair type=vc;  
run;
```

Model 3

```
proc mixed data=mztwin12 method = ml covtest noclprint;  
class pair;  
model f11 = religdev religavg solution;  
random intercept subject=pair type=vc;  
run;
```

Model 4

```
proc mixed data=twin12 method = ml covtest noclprint;  
class pair MZ twinid;  
model f11 = religdev religavg religdev*MZ solution;
```

```

random intercept subject=pair group=MZ type=vc;
repeated / group=MZ type=vc;
run;

```

Sas Output For Model 4

The SAS System 12:43 Thursday, December 1, 2011 75

```

The Mixed Procedure
Model Information
Data Set WORK.TWIN12
Dependent Variable f11
Covariance Structure Variance
Components
Subject Effect PAIR
Group Effects MZ, MZ
Estimation Method ML
Residual Variance Method None
Fixed Effects SE Method Model-Based
Degrees of Freedom Method Containment
Dimensions
Covariance Parameters
4
Columns in X
5
Columns in Z Per Subject
2
Subjects
644
Max Obs Per Subject
2
Number of Observations
Number of Observations Read
1370
Number of Observations Used
1286
Number of Observations Not Used
84
Iteration History
Iteration Evaluations -2 Log Like
Criterion
0 1 3035.96177444
1 2 2906.05165136
0.00000000
Convergence criteria met.
Covariance Parameter Estimates
Standard

```

Z

Cov Parm Subject Group Estimate Error

Value Pr > Z

Intercept PAIR MZ 0 0.2112 0.03214

(CONTINUED ON NEXT PAGE)

6.57 <.0001

Intercept PAIR MZ 1 0.3354 0.04633

7.24 <.0001

Residual MZ 0 0.3889 0.02778

14.00 <.0001

Residual MZ 1 0.3172 0.02831

11.20 <.000

The SAS System 12:43 Thursday, December 1, 2011 76

The Mixed Procedure

Fit Statistics

-2 Log Likelihood

2906.1

AIC (smaller is better)

2922.1

AICC (smaller is better)

2922.2

BIC (smaller is better)

2957.8

Solution for Fixed Effects

Standard

Effect MZ Estimate Error DF

t Value Pr > |t|

Intercept -0.4036 0.06920 641

-5.83 <.0001

religdev -0.00848 0.01681 641

-0.50 0.6138

religavg 0.05042 0.006817 641

7.40 <.0001

religdev*MZ 0 0.02717 0.02068 641

1.31 0.1894

religdev*MZ 1 0 . .

. .

Type 3 Tests of Fixed Effects

Num Den

Effect DF DF F Value

Pr > F

religdev 1 641 0.24

0.6221

religavg 1 641 54.69

<.0001

religdev*MZ 1 641 1.73

0.189

Appendix B: Mplus Code

```

data: file = model6.txt;
variable:
names = pair zygo f11a f11b religa religb;
missing =.;
grouping = zygo (1=mz 2=dz);
usevariable = religa religb f11a f11b;
analysis:
type = missing h1;
model = nocovariances;
model:
a11 by religa@1;
c11 by religa@1;
e11 by religa@1;
a21 by religb@1;
c21 by religb@1;
e21 by religb@1;
[religa* religb*] (relmean);
[a11-e21@0];
a11*4 (amz);
c11*8 (c);
e11*4 (e);
a21*4 (amz);
c21*8 (c);
e21*4 (e);
a11 with a21*4 (amz);
c11 with c21*8 (c);
religa@0;
religb@0;
religa with religb@0;
f11a on a11*.01 (areg)
c11*.04 (creg)
e11*.01 (ereg);
f11b on a21*.01 (areg)
c21*.04 (creg)
e21*.01 (ereg);
a12 by f11a@1;
c12 by f11a@1;
e12 by f11a@1;
a22 by f11b@1;
c22 by f11b@1;
e22 by f11b@1;
[f11a* f11b*] (delmean);
[a12-e22@0];
a12* (xmz);

```

```
c12* (y);
e12* (z);
a22* (xmz);
c22* (y);
e22* (z);
a12 with a22* (xmz);
c12 with c22* (y);
f11a@0;
f11b@0;
f11a with f11b@0;
model dz:
a11 with a21*4 (adz);
a12 with a22* (xdz);
model constraint:
amz= 2*adz;
xmz =2*xdz;
output:
sampstat tech1 cinterval standardized;
```


Chapter nine Methods of Small Group Research

Norbert L. Kerr and R. Scott Tindale

This chapter seeks to inform the reader about how research on group process and outcomes is conducted. But before turning to these topics, we thought that it useful to describe just what such research actually studies. The word “group” has a time-honored place in social psychology (Forsyth, 2010). However, as with many terms with a long history in the field, this word has been used in a number of different ways over the years. For instance, the term has often been used – particularly by scholars of stereotyping and intergroup relations – to refer to any aggregate of people who share some socially salient characteristic(s) – for example, a racial, ethnic, gender, or national “group.” In this chapter, however, “group” refers to something different and quite distinct – a type of social entity that in the literature often has been called the *small group* (e.g., Hare, 1976; Haythorn, 1953). More specifically, the *small group* refers to a collective of persons whose history of shared fate, common purpose, and interaction has led to the perception, by participants and outsiders alike, that this collective is a social unit (Campbell, 1958; Heider, 1958).¹ We view the idea of common purpose – particularly as it involves coordinated task activity – as the essential feature that distinguishes the small group from other types of social units (e.g., close relationships; cf. Weber & Harvey, 1994).

Moreover, many phenomena that occur in small groups also occur in situations that do not involve a real social entity; rather, they occur in settings in which participants (temporarily) work together to accomplish some goal(s) with few, if any, feelings of “groupness.” We will refer to the inclusive set of contexts – including both small groups (as defined earlier) and temporary, task-oriented collectives – as *group contexts*. A broad concern with group contexts rather than more narrowly on small groups per se can be justified for many reasons, not the least of which is that most investigations of group process and outcomes have studied these issues by examining people in temporary group contexts rather than actual small groups.

The enduring and often indeterminate time frame of “real” groups, to say nothing of their inherent complexities, makes their systematic study a daunting

enterprise. And even the study of collective activities in more easily structured group contexts can be challenging enough, given the complicated phenomena of interest. What are those phenomena? The topics that we present in [Table 9.1](#) reflect the primary questions addressed in classic and contemporary research on group functioning (Forsyth, 2010; Levine & Moreland, 1990, 1997; Wheelan, 1994). Students who are drawn to the complex problems of individuals interacting in groups often ask, as they consider committing themselves to such a labor-intensive enterprise: What questions are so special to this field that it is worth expending the great effort needed to answer them? What can be learned that can justify investments of such magnitude? In this chapter we also attempt to address these questions, to explain why the exploration of people's behavior in group contexts is a critical task for social psychology. In doing so, we argue that the phenomena are unique, the methods robust, and the outcomes of great importance to social psychology. The pages that follow, then, attempt to explore contemporary methods for conducting research on group phenomena and to convince the reader that investigating something as complex as individual behavior in groups can be stimulating and rewarding.

TABLE 9.1. Major Topics, Paradigms, and Variables of Group Research

Substantive Topic/Area and Core Questions	Representative Paradigms (and Articles)	Representative Independent Variables	Representative Dependent Variables
Intragroup processes			
Group formation and development			
What functions does group membership serve?	Festinger's cohesiveness paradigm (p. 238) (Back, 1951)	Relevance of task to the group	Level of group cohesiveness
How are group members recruited and socialized?	Newcomb's acquaintance-process paradigm (p. 188) (Newcomb, 1961)	Other members' resources and knowledge	Distribution of speech acts
Do groups go through standard phases of development or work?	The affiliation paradigm (p. 193) (Schachter, 1959)	Task type	Desire to affiliate
	Levine and Moreland's newcomer paradigm (Moreland, 1985)	Group size	
Group structure			
What is the pattern of relationships (liking, power, status, communication, etc.) among group members?	Schachter's productivity-norm paradigm (p. 123) (Schachter et al., 1951)	Task features	Task performance
What is the effect of such patterns on group functioning?	Adam's inequity paradigm (p. 204) (Walster, Walster, & Bersheid, 1978)	Allowed patterns of communication	Evaluation of group members
What expectations of member behavior (e.g., roles and norms) develop and guide behavior in the group?		Group cohesion	Allocations to self vs. others
			Perceived social norms/role
Group Communication			
Who says what to whom?	Bales's Interaction Process Analysis (IPA) paradigm (p. 142) (Bales & Strodtbeck, 1951)	Size of group	Participation rates
How are member characteristics related to amount and type of communication?	The Communication Network paradigm (p. 168) (Leavitt, 1951)	Type of problem	Distribution of comments within a coding scheme
How does the amount and type of communication affect one's status in the group?	The Valence Coding paradigm (Hoffman & Maier, 1964)	Permitted communication links	Subjective ratings of influence or leadership
Can group preference or solution be predicted from patterns in the content of communication?	The Hidden Profile paradigm (Stasser & Titus, 1985)	Distribution of shared and unshared information in the group	Solution quality
How efficiently do groups elicit task-relevant information from their members?			

(continued)

Substantive Topic/Area and Core Questions	Representative Paradigms (and Articles)	Representative Independent Variables	Representative Dependent Variables
Social influence processes What are the basic processes through which group members exert influence on one another? What personal and situational factors lead to leadership emergence and effectiveness?	Asch's conformity paradigm (p. 235) (Asch, 1951) Sherif's group norm paradigm (p. 234) (Sherif, 1936) Milgram's obedience paradigm (p. 181) (Milgram, 1974) Bystander-intervention paradigm (p. 231) (Latané, & Darley 1970) Social-learning paradigm (p. 230) (Bandura, 1962) Reaction to deviate paradigm (p. 239) (Schachter, 1951) The leader style paradigm (p. 255) (Lewin, Lippett, & White, 1939)	Task type Level of group cohesiveness Levels of power/status of influencer Relationships between members Leadership styles	Level of compliance Imitative behavior Inclusion/exclusion from the group Group performance
Group productivity How do member, group, and task features affect group productivity? What factors affect whether groups achieve, fall short of, or even exceed their nominal potential productivity?	Social-facilitation paradigm (p. 228) (Zajonc, 1965) Laughlin's concept-attainment paradigm (p. 70) (e.g., Laughlin & Johnson, 1966) Participatory decision making paradigm (p. 123) (Coch & French, 1948) Social loafing paradigm (Latané, Williams, & Harkins, 1979)	Presence of others Distribution of member abilities, personalities, etc. Group size	Task performance Member arousal Member contributions
Group decision making Are there systematic rules linking individual and group choices? Under what conditions are group decisions of higher or lower quality than individual decisions? What unique processes distinguish group from individual decision-making processes?	Lewin's group discussion paradigm (p. 232) (Lewin, 1953) The Risky-shift paradigm (p. 81) (Wallack, Kogan, & Bem, 1962) Davis' mock-jury, SDS paradigm (p. 85) (Davis, Kerr, Atkin, Holt, & Meek, 1975) Groupthink paradigm (Janis, 1982) Collective induction paradigm (Laughlin, 1996)	Public vs. private discussion Type of decision task Procedural factors Group composition	Fulfilling intentions expressed in groups Contrast of individual and group judgment Distribution of group decisions Functional relation between individual and group decisions (social decision scheme)
Intragroup conflict How do patterns of group member interdependence guide member behavior? What are the ways members exchange resources to resolve such conflicts (e.g., through bargaining, negotiation, coalition formation)? How do group members reconcile conflicts between personal and collective interest?	The prisoner's/social dilemma paradigm (p. 103) (Rapoport, 1976; Brewer & Kramer, 1986; Dawes, McTavish, & Shaklee, 1977) The bargaining paradigm (p. 99) (Siegal & Fouraker, 1960) Deutsch's Trucking game (p. 106) (Deutsch & Krauss, 1962) The Coalition paradigm (p. 110) (Komorita & Chertkoff, 1973)	Game/task features Prior training and experience Social motives	Absolute and relative gain of group members Levels of cooperation and competition
Environmental processes How do features of the physical environment affect group and group-member behavior? How do groups regulate their use of physical environments Extra-group processes	The Westgate-Westgate West paradigm (Festinger, Schachter, & Back, 1950) Sommer's personal space paradigm (p. 217) (Sommers, 1959) Groups-in-isolation paradigm (p. 218) (Altman & Haythorn, 1967) Crowding-performance paradigm (Freedman, Klevansky, & Ehrlich, 1971)	Functional distance between group members Seating positions Temporal demand	Territorial behavior Task performance Interpersonal attraction/hostility
Groups as contexts for action How does being in a group, particularly in a very large groups or crowd, alter thinking and action?	The deindividuation paradigm (Diener, Lusk, DeFour, & Flax, 1980) Kelley's emergency-escape paradigm (Kelley, Condry Dahlke, & Hill, 1965)	Group/crowd size Level of anonymity	Antisocial behavior Counternormative behavior
Intergroup relations What are the causes and cures of intergroup conflict? How does group membership alter social perception?	Sherif's Robber's Cave paradigm (p. 118) (Sherif et al., 1961) The minimal-group paradigm (Tajfel, Billig, Bundy, & Flament, 1971) The in-/outgroup homogeneity paradigm (Judd & Park, 1988)	Group membership Permeability of group boundaries Level of intergroup conflict of interest	Intergroup conflict Allocation of resources to in/outgroup members Perception/evaluation of in/outgroup members

Note: All page references enclosed in parentheses refer to McGrath (1984).

Why study groups? When you watch people in their natural habitat, it is clear that the small human group is a (perhaps, the) primary unit of social psychology.

Ordinary human behavior, which can be observed on any street corner, occurs between people who live within groups and who go between groups. In their ongoing behavior, people affect each other in ways that cannot be sufficiently explained by knowledge of the attributes of the individual actors. Groups are one of the primary devices human beings have to accomplish their purposes. What better for a social psychologist to study?

Before turning to the real substance of this chapter,² we want to offer a less glib answer to this important question. One common and reasonable answer is that group phenomena (defined restrictively or not) are ubiquitous. We will never have a comprehensive understanding of human social behavior without an understanding of human social groups. This proposition probably would not be very controversial among social psychologists, yet even though practically every social psychologist would say that what he or she studies is highly relevant to a full understanding of behavior in groups, only a minority of our discipline would say they study group phenomena. What distinguishes this remnant of what was once a thriving enterprise in social psychology (cf. McGrath & Altman, 1966; Steiner, 1974) from the currently more dominant individualistic-cognitive paradigm (Moreland, Hogg, & Hains, 1994; Steiner, 1986)? One thing is a conviction on the part of group researchers that we shall come to that universally desired understanding of group behavior faster and more deeply by focusing our attention on behavioral settings that have certain properties, properties that we might term the four I's: interaction, interdependence, identification (with something bigger, more inclusive than the self), and imbeddedness (in interpersonal social structures, such as role structures, power relationships, normative systems, etc.).

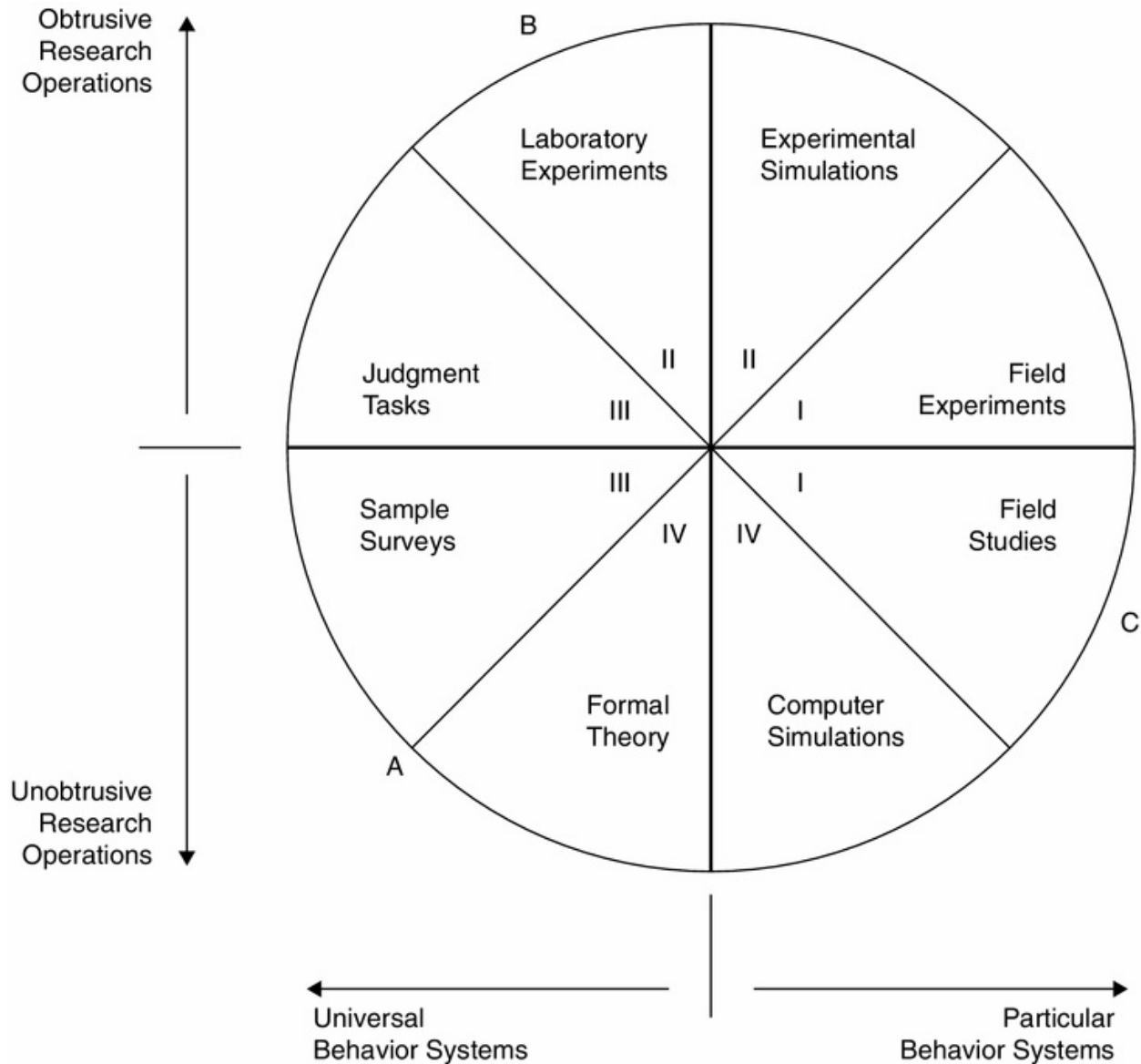
Implicit, we think, in the working assumptions of most small group researchers is the conviction that it is not always productive to analyze phenomena at the most molecular level possible, and that some issues are better, more insightfully addressed at a more molar level of analysis (see Steiner, 1974, 1986). For example, it is possible, in principle, to describe the “behavior” of the helium in a balloon as the net effect of the movements of billions of individual helium molecules. Such an approach might depend on describing the “actions” and interactions of individual molecules and would, of necessity, result in enormously complex descriptive or explanatory models. However, the basic laws of thermodynamics turn out to offer simple relationships between certain summaries of the behavior of those billions of individual molecules – such as the temperature, volume, and pressure of the gas – which are much more useful for most purposes than (literally) more molecular models. Likewise, group

researchers assume that there will be times when concepts defined at the group level may be more powerful or efficient for advancing our understanding of behavior than concepts defined at more molecular (e.g., individual) levels. (A similar presumption pervades all of social psychology – we take for granted that analyses of social behavior undertaken at the level of the individual can often be more useful or tractable than analyses at more molecular levels [e.g., physiological, neuronal, cellular, genetic].) This is not just an article of faith; there are many good illustrations in the social-behavioral sciences of the greater utility of molar analytic approaches. For instance, it has been hypothesized (Steiner, 1972) and shown (e.g., Hill, 1982, for a review) that task groups usually fall short of their productive potential. Bray, Kerr, and Atkin (1978), for example, showed that for a certain kind of intellectual task, this suboptimality increased as groups became larger. Now, this phenomenon could be analyzed at the individual level, in terms of the effects of increasing group size on the different perceptions and actions of individual group members. But a simple and efficient understanding of the full pattern of data results from the use of a group-level concept (viz., the group's functional size, which is that group size \tilde{n} whose productivity matches the observed productivity of the n -person group). In particular, for simple intellectual problems, Bray and colleagues found that \tilde{n} was 1 (or sometimes as much as 2), no matter how large the group actually is. That is, when participants take turns talking about such problems in a face-to-face group, the group ends up functioning about as well as would be expected if there were only one person in the group (cf. Diehl & Stroebe, 1987). Although one could probably also describe this phenomenon by reference to individual perception (e.g., perceived competition for speaking time, felt individual responsibility), in terms of predicting and understanding group performance, little may be gained in doing so.

Generic Strategies for Small Group Research

It is well recognized that any single study can, at best, test only some aspects of a proposition (cf. Brewer & Crano, Chapter 2 in this volume), much less test all aspects of all related propositions. In research, including group research, students need to recognize that not only are multiple studies needed to confirm a hypothesis but also that entirely different methods may be needed as well. Different methods are required to compensate for the inherent weaknesses in any particular choice of method. Runkle and McGrath (1972) have developed this argument systematically in their circumplex model of research methodologies.

They identify eight generic research methodologies that they array like pieces of a pie (see [Figure 9.1](#)). Three points on the circumference of the circumplex (marked A, B, and C in the figure) mark points of maximum concern with, respectively, generality over actors (presumed in most surveys and formal theories), precision of measurement and control (maximized in laboratory experiments), and preserving the naturalism (“system character”) of some particular context (maximized in field studies). By imposing this spatial representation of methods, Runkle and McGrath underscore the important point that there is no single best method of inquiry – each method has its inherent strengths and weaknesses, and one can never simultaneously enjoy the former and avoid the latter. In choosing a method, one is perpetually on the horns of a dilemma. And it is an illusion to believe that one can, like some nimble matador, so shuffle or position oneself that one is never caught on at least one of the horns of that dilemma. For example, most social psychologists, by dint of their training and adherence to professional norms, opt to avoid at all costs the horn of internal invalidity, and prefer to stick to methods near point B of the circumplex, but by so doing they insure that their research will be far from both point C (and hence highly artificial) and point A (and hence likely to characterize a very particular population – typically, the college sophomore). The only solution for these dilemmas, of course, is to employ multiple methods across investigations and hope that their findings will converge on some conclusions that transcend each separate method's limitations (Brewer & Crano, Chapter 2 in this volume; Campbell, [1969](#)).



- I. Settings in natural systems.
- II. Contrived and created settings.
- III. Behavior not setting dependent.
- IV. No observation of behavior required.
- A. Point of maximum concern with generality over actors.
- B. Point of maximum concern with precision of measurement of behavior.
- C. Point of maximum concern with system character of context.

Figure 9.1. Runkle & McGrath's Method Circumplex.

In what follows we elaborate on these themes, using the general structure of the circumplex to focus on the use of several generic research strategies for the

study of small groups, noting some of the distinctive tools, challenges, and limitations associated with each.

Field and Archival Research on Groups

Group processes and outcomes can be, and often have been, studied outside the laboratory using nonexperimental methods (e.g., Aronoff, 1967; Moreno, 1953; Roethlisberger & Dickson, 1939; Whyte, 1943), and a good portion of this work has investigated actual small groups. Such field and archival research has a rich tradition in social science, generally, but is not frequently conducted in contemporary social psychology (e.g., Reis & Stiller, 1992). Moreover, a detailed exploration of these approaches would require much more space than we can devote in a single chapter on group process research. Thus, in this section, we provide only a basic overview of these methods as they have been applied to the study of group phenomena, primarily by citing some representative examples from the literature. Other sources discuss these techniques more comprehensively (e.g., Judd, Smith, & Kidder, 1991; Weick, 1985).

Observational Field Methods

Much can be learned about group processes and outcomes – as well as a host of social phenomena in general – by carefully observing people's everyday (and not-so-everyday) experiences as they occur. The methods available to study group phenomena in field settings include the usual variations of observation and interview (also see Heyman, Lorber, Eddy, & West, Chapter 14 in this volume; Reis, Gable, & Maniaci, Chapter 15 in this volume). For example, Muzafer Sherif, one of the founding fathers of social psychology, studied the evolution of group structure, entitativity, cohesiveness, and actual intergroup conflict by observing the activities of participants at a boys' summer camp (e.g., Sherif, Harvey, White, Hood, & Sherif, 1961). In this context, subsets of campers (who had never previously met) were formed into aggregates as a function of cabin assignment, given group names (e.g., "Red Devils," "Bull Dogs"), and assigned to perform a number of activities (e.g., preparing a cookout meal, practicing baseball as a team, etc.). Although participants were informally interviewed periodically, the bulk of the data that Sherif collected was derived from careful observations that he and his staff made of the campers' activities. For instance, Sherif gained understanding into emergent social structure by observing how the boys acted with regard to one another as they went about

performing tasks. Here is how Sherif (1966) described a cookout:

The staff supplied the boys with unprepared food. When they got hungry, one boy started to build a fire, asking for help in getting wood. Another attacked the raw hamburger to make patties....A low-ranking member took a knife and started toward the melon. Some of the others protested. The most highly regarded boy in the group took the knife, saying, "You guys who yell the loudest get yours last."

(p. 77)

These and other observations yielded many useful insights into group development and functioning.

The distinctive strength (cf. Runkel & McGrath, 1972) of a field study is its naturalness; one can examine behaviors of interest as they naturally occur. Field studies – such as Sherif's (1966) classic work – ideally exploit this strength. One common purpose is to discover natural phenomena that need to be understood. Many of the classic topics in social psychology (rumor transmission, opinion change, organizational effectiveness, obedience, conformity, helping, attraction, prejudice, etc.) began with a special experience or arresting observation of some aspect of ordinary, "real" life. Another common purpose of a field study is to confirm that our knowledge of those phenomena – based largely on more controlled research methods used in settings that are necessarily more artificial – generalizes to natural behavioral settings. Field studies can be difficult, expensive, and tedious, but no other method can better establish whether a social process is important, in terms of its actual effects in real social settings, what range of factors need to be examined, and its full network of associations with other social factors (Reis, 1983).

The weaknesses of studying group phenomena in this way are as clear as its strengths. Beyond certain potential biases discussed later (e.g., bias that can result when an outsider intrudes on a natural group's functioning, or when the author of a hypothesis is directly involved in data collection), research hypotheses are usually causal but the data in a field study are, at best, correlational. The variables being observed may well be markers for quite different, but even more important uncontrolled, unmeasured, and confounding ones. There is often no way to know. In principle, one might be able to resolve such ambiguities by additional measurement or manipulation, but this possibility requires one to have some control of the phenomena in question, and the essence of the natural field setting is that events are controlled by natural processes, not

by the investigator.

Traditionally, observational field methods have been divided into two principal types: those in which the researcher strictly maintains his bystander status as events unfold (*nonparticipant observation*) and those in which the researcher, at least to some extent, participates in the activities of interest (*participant observation*). Both types are used to study group processes and outcomes, so each is briefly discussed in the sections that follow.

Nonparticipant Observation

The “Bank Wiring Room” Study, which was part of one of the first attempts by behavioral scientists to systematically study the industrial workplace, is a classic example of nonparticipant observational field research on group phenomena (Mayo, 1933; Roethlisberger & Dickson, 1939). For this study, researchers received permission from a large telephone equipment manufacturing company to relocate a work group, whose job it was to produce banks of electrical switches, to a smaller room that was off to the side of the main plant area. A member of the research team sat at a desk off to the side for the many weeks that the group used this room. This person was basically “a fly on the wall,” who observed and recorded what the group members did. Some of the data that the observer recorded were specific regular activities (e.g., who initiated interactions with whom), whereas others were summaries of more singular events (e.g., an incident in which one person ventured into the main plant to fetch supplies).

These records were handwritten – an arduous and labor-intensive task – and the researcher was often required to both observe and record at the same time. However, there have been substantial advances in recording technology since the time of this classic study. Contemporary research of this type would utilize digital videorecording equipment to collect data. Among the manifest advantages are: (a) Videorecording yields records of what has transpired that are verbatim, rich in detail, and permanent. As such, researchers do not have to decide what is important to observe before the events in question take place. They can review the recordings over and over again before deciding what data should be distilled. (b) Data distillation itself is less stressful and potentially much more accurate from videorecordings than from coding “online.” Judges and coders who work with recordings essentially are nonparticipant observers with two major advantages: They can “collect” data at their own pace, rather than be forced to record at the speed with which events are unfolding; and they

can use the rewind button to reexamine ambiguous behavior. (c) The miniaturization of videorecording equipment now permits a camera to be truly unobtrusive.

The truly raw-data nature of video observational records can also be a major disadvantage. Videorecordings capture everything that the camera “witnesses,” for as long as the camera is operating. Recording all the time that the work group spent in the bank wiring room, for instance, would have used a massive amount of memory. It would have been a daunting task just to have coders view the recordings to edit out unnecessary footage. Additionally, coding recordings for particular events of interest, whether from videorecordings or as the events occur, requires a number of strategic methodological choices (McGrath & Altermatt, 2001).

Of course, researchers can opt to time-sample the events of interest (see Heyman et al., Chapter 14 in this volume), but this solution also has potential problems. Because the equipment lacks the capacity to judge when to record, the researchers must make that decision. Employing some sort of a priori, intermittent, fixed, or variable sampling scheme leaves open the possibility that an important incident will be missed. Another approach is to have a researcher present at all times during observation periods to make moment-by-moment decisions about what should be recorded. This is pertinent when a discrete event is of interest (e.g., a particularly important decision in a group discussion). However, sampling is optimal when an extensive record has been obtained and the relative frequency of different “kinds” of behavior (e.g., leadership behavior) needs to be obtained across all members of a group.

From the foregoing discussion, it should be clear that there are no simple criteria for deciding whether to observe and record online or use recording equipment to produce verbatim accounts for later use. As with much of the research process generally, such decisions have to be made by informed researchers who understand both their particular circumstances and various advantages and disadvantages of each approach.³

Participant Observation

As noted, field researchers sometimes “observe from within” by becoming actual participants in a group's experiences. Historically, participant observation has been used much less frequently in social psychology than in other social sciences, particularly anthropology and sociology, but there are a few instances

of its use in our discipline. One noteworthy example (Festinger, Reicken, & Schachter, 1956) involved participant observation of a very unusual group, whose task was to make sure that some humans survived a prophesized destruction of the world. Festinger and his colleagues watched the unfolding events from “the inside.” Even though they attempted to maintain a low profile and not do anything that would affect what was transpiring, the researchers still had to “behave normally” as group members; as such they took part in the group's activities and behaved in much the same way as everyone else. (Needless to say, the moment of reckoning did come and go as Festinger and colleagues had hoped, and the investigators were able to make interesting observations of what happens psychologically when prophecy fails.)

The obvious advantage of this approach is that it provides the researcher with a unique opportunity to observe particular group processes and outcomes firsthand and in situ. In this way she or he has the potential to learn about phenomena of interest that are unavailable to external observers. The major disadvantages concern measurement. To some researchers this method can rarely be scientific because observations are usually impressionistic and nonsystematic. A related problem involves potential reactivity. Ideally, participant observers act in ways that have no impact on the phenomena of interest. But behaving with complete neutrality is no easy feat, and because there typically is no way to verify that the researcher's presence, appearance, and actions did not influence events, the naturalness that is the distinctive advantage of all field methods may be compromised. Finally, participant observation also tends to be very time consuming and costly.

Archival Studies

There is a wealth of underutilized archives of many different kinds available to test our hypotheses, longitudinally, cross-culturally, or within any particular culture. Such archives may have been explicitly created and maintained for research purposes or represent records collected for other purposes altogether (e.g., the U.S. Census, newspapers, organizational records). The data may be suitably recorded for direct analysis or may require considerable sifting and recoding. There are several clear advantages and disadvantages of archival research. One of the clear advantages is that the data has already been collected; this can be a significant advantage in research on groups, given the extra time, effort, and cost it routinely entails. And because someone other than the investigator has collected the data, the risk of experimenter expectancy effects is

reduced. In some instances, an archive's data can also be much more voluminous and varied than might be possible through planned, direct observation. Archival research can also often simplify matters of institutional review for participant protection, particularly when the records are public or the original participants had already given permission for their behavior to be recorded. Archives can also permit examination of questions that might be unfeasible to address otherwise. For example, studies of life span development (including the development of long-standing groups) or even longer historical comparisons (jury composition in early vs. contemporary American history) may require archival data. Or events that are unpredictable or infrequent may be easier to locate in archives than to await and observe. Most of the disadvantages stem from the fact that the investigator usually has little or no control over what has been recorded or how it has been recorded. This may mean that important observations may be missing or retrievable only through labor-intensive search and coding, measurement criteria may have changed across time, or the reliability of measurement may be low or indeterminable.

An archival approach to hypothesis testing may be illustrated by Tetlock's (1979) investigation of Janis's groupthink hypothesis. The public statements made by key decision makers (presidents, secretaries of state) in five U.S. foreign policy crises were coded for the integrative complexity of the decision makers' thinking and their positivity/negativity toward in-group/out-group symbols. Three of these crises had previously been identified by Janis as exemplars of groupthink (e.g., the invasion of North Korea); the other two exemplified well-formulated, vigilant decision making, where groupthink was avoided (e.g., the Cuban Missile crisis). Tetlock was able to confirm certain groupthink predictions (e.g., leaders were more simplistic in their thinking in the groupthink crises), but got less support for others (e.g., leaders were not more negative toward out-group symbols in the groupthink crises).

Field Experiments

A field experiment introduces direct manipulation of some variable of interest within a field setting. This method can combine the strengths of a field study with the distinctive strengths of an experiment – the ability to draw causal inferences. However, as Runkle and McGrath (1972) caution us, by imposing some degree of control over context and measurement, one inevitably makes the research setting less natural than a field study is, while never achieving the high degree of control of a lab experiment. Field experiments are rarely undertaken

because having all the necessary elements in place at the right time can require special access to and control of field settings. Such control can be difficult to acquire and maintain, particularly when experimental requirements (e.g., random assignment, intrusive measurement) interferes with the usual operation of the setting, or when the results of the field experiment might threaten the norms, status, or even continued existence of the groups or organizations being studied.

One nice illustration of a field experiment on small groups is Hannaford, Hans, and Munsterman's (2000) study of the effects of predeliberation discussion of a case among civil jury members. As part of a review of jury procedures, the state of Arizona considered several innovations, including allowing jurors to discuss the trial evidence prior to their formal deliberations. Armed with an Arizona Supreme Court administrative order permitting trial judges to depart from the usual instructions (prohibiting any predeliberation discussion), the authors were able to get trial judges to give instructions that permitted predeliberation discussion to 84 randomly selected civil juries and traditional, no-discussion instructions to 73 other juries. With the assistance of the court administrators, they not only were able to collect publicly available outcomes (e.g., verdicts, awards), but were able to get attorneys, judges, and jurors to fill out questionnaires probing their reactions. A number of interesting findings emerged: Juries that could discuss the case were more certain of their preferences prior to deliberation and were less likely to reach unanimous agreement. For present purposes, equally interesting were some of the methodological ambiguities that arose from doing a field experiment. For example, even though cases were purportedly assigned to condition randomly, systematic differences in cases emerged (e.g., cases assigned to the No Discussion condition were rated by the judges as significantly more complex than those assigned to the Discussion condition). This could be attributable to chance but could also reflect hard-to-detect departures by court personnel from strict random assignment. And court procedures in this real-world context meant that the manipulation could only vary jurors' *permission* to discuss the case. As it turned out, a substantial fraction (31%) of those juries that could discuss the evidence never did so. This is much like a clinical drug trial where one could be misled about the effectiveness of a drug if a third of those in the drug-treatment group failed to take the medication. And because it was not possible to know in advance which trials would be in each condition, it was not possible to test jurors' memory of trial content, which has been alleged to be improved via jurors' discussions. In short, lack of control and opportunity for measurement necessarily limited this field experiment's internal validity and scope.

Experimental Methods

All the methods “above the equator” in Runkle and McGrath's circumplex (see [Figure 9.1](#)) could be classified as experimental methods. To varying degrees, they all strive to emulate the idealized “true experiment” (Anderson, 1966), which manipulates one or more potential causal variables, controls all other variables, and measures one or more dependent variables of interest (Smith, Chapter 3 in this volume). As noted earlier, field experiments sacrifice a good deal of control to preserve greater naturalness of context; experimental simulations try to retain certain essentials of the natural context of interest while gaining even more control and opportunity for observation. The laboratory experiment generally achieves maximal control and observation opportunity, but can, at most, focus on a generic or abstract set of natural contexts of interest. In judgment tasks, there is even less concern with the fidelity of context, but maximum concern with how carefully chosen and presented stimuli are judged. Within research on the psychology of juries, for example, this spectrum of methods is illustrated by (1) field experiments like Hannaford et al.'s (2000); (2) jury simulation studies, which strive for fidelity to the essence of the jury's task and courtroom context (cf. Kerr & Bray, 2005); (3) highly controlled lab studies of social influence in groups seeking consensus on an arbitrary issue (e.g., Godwin & Restle, 1974); and (4) a study of what features of a human face make it memorable (e.g., to an eyewitness of a crime; e.g., Chance, Goldstein, & McBride, 1975).

Although a well-controlled experiment cannot provide confidence that a phenomenon is important (in any real-world setting of interest), robust, or widely relevant to aspects of the larger society, it nevertheless provides the best method that we have to get a reasonable grasp on the causal antecedents of a social process (see Brewer & Crano, Chapter 2 in this volume). These virtues have led to this becoming the preferred method for social psychological inquiry (Rozin, 2001; Sears, 1986), including inquiry on group behavior.

We have mentioned that experimentation on groups entails a number of unique costs, compared to experimentation on individuals. The most obvious cost is that of obtaining n participants for every replicate in a study of n -person groups. Some studies, like Kerr & MacCoun's (1985) experimental comparison of 3-, 6-, and 12-person mock juries, can require very large participant pools. Besides large pools, deep pockets, and persistence, there are a few other ways of reducing such costs. For example, one can minimize wasted sessions (because of too few participants) or wasted participants (when more show up than are

required) by over-scheduling and running multiple groups at each experimental session. Of course, this can also require more experimenters and lab space per session. The possibility of distributed or virtual groups, discussed later in the chapter, may offer one means of overcoming some of the logistical problems associated with scheduling face-to-face groups.

A related difficulty arises when one wishes to compare groups with, among other things, particular compositions of ability (e.g., Laughlin, Branch, & Johnson, 1969), attitudes (e.g., Anderson, 1975), personality (e.g., Lampkin, 1972), or gender (e.g., Kent & McGrath, 1969). Again, composing many groups from a large and diverse set of participants is most efficient in such cases.

Systematic Observation of Groups

Many theories and frameworks underlying research on small groups imply that group process is a key component of group outcomes (Hackman & Morris, 1975; McGrath, 1984). Although it is usually straightforward to assess outcomes, assessing group processes can be much more difficult (Weingart, 1997). In many instances, group processes are either inferred by the outcomes (e.g., good outcomes stem from good processes) or are assessed retrospectively through questionnaires. Retrospective reports can be useful and in some settings may be the only means available for studying group process. However, with advances in both theoretical precision and technological sophistication, greater emphasis has been placed on assessing process through systematic observation and analysis of actual group interaction (although, given its labor intensiveness, such analyses are still the exception rather than the rule; Moreland, Fetterman, Flagg, & Swanenburg, 2010).

Two rather different approaches toward measuring group process have been prevalent in the literature (Weingart, 1997). The first involves developing a scheme for coding group interaction that will work in almost any small group context (Bales, 1950; Futoran, Kelly, & McGrath, 1989), whereas the second attempts to design the scheme around the specific task of interest (e.g., Hastie, Penrod, & Pennington, 1983; Weldon & Weingart, 1993). A fairly recent example of this first type was developed by Futoran *et al.* (1989) and called TEMPO (Time by Event by Member Pattern Observation system). The system attempts to combine aspects of activity-based coding systems (those looking at who talked to whom with what frequency (see Stasser & Taylor, 1991) and more process-oriented schemes (e.g., IPA system; Bales, 1950). Thus, units of time are coded for instances of various different types of acts or behaviors. Each act is

assigned to a specific member and a function category. The function categories fall into two broad classes: content and process. Within each class, acts are coded as either proposals or evaluations. Content statements refer to task-relevant ideas or concerns, whereas process statements refer to goals or strategies associated with carrying out the group task. A series of non-task-related categories are also defined (see Futoran et al., 1989 for a more complete description). The strengths of the system include its focus on time and temporal contingencies, comprehensiveness, and appropriateness for virtually any type of task-oriented group.

Among the many task/situation-specific group interaction coding schemes, a particularly nice example is the one Hastie, Penrod, and Pennington (1983) developed for studying jury deliberation. Because the purpose of their study was to assess jury performance, they designed the process measures around five performance criteria that well-performing juries should meet: juries should provide a representative cross-section of the population; they should express a variety of perspectives; they should be accurate fact finders; they should accurately follow the pertinent law; and they should reach an accurate verdict. All mock juries saw the same trial, so one of the coding schemes focused on whether key pieces of evidence were recalled and discussed. Hastie and colleagues also coded the videorecordings of jury deliberation for accurate and inaccurate mentions of the judge's instructions and key aspects of the verdict definitions. A third coding scheme took a more functional view (like TEMPO) and coded statements as questions, suggestions, and so on, but with some categories being specific to jury discussions (suggested verdicts, corrections to misstated evidence, etc.). In addition to coding statements into categories, Hastie and colleagues also looked at the process from three additional perspectives. First, they looked at participation rates by juror and by verdict preference in order to assess whether different perspectives were given equal time. Second, they looked at deliberation time as another aspect of process, not only in terms of overall deliberation time but also time associated with different types of deliberation content and at what point in time certain types of statements were made (e.g., when legal issues were discussed vs. evidence in terms of the deliberation sequence). Finally, they tracked influence processes in the juries by estimating transition probabilities for groups moving from one particular verdict distribution to another (see Kerr, 1981).

Both generic process measures and more tailored versions have their benefits and costs. More general schemes can be used to compare groups working on different types of tasks and can also be used to track changes in processes over

time as groups move from one task to the next. They may also come with training manuals so researchers do not have to “reinvent the wheel” for each new attempt at measuring group process. However, their generality also impedes their usefulness for assessing the importance of task-specific content and processes. As was evident in the Hastie *et al.* (1983) example, even systems designed for a specific type of group often borrow from general schemes that have proved useful in the past. Thus, most instances of group interaction analysis tend to use a combination of general systems with adaptations to the current task and group environment.

Although there is no single best way to study group process, McGrath and Altermatt (2001) provide six partially conflicting rules that researchers would be wise to consider when thinking about studying group processes. First, they suggest researchers plan ahead to make sure that their coding scheme or assessment procedure can capture the aspects of process they believe will be important. Thus, planning based on previous theory and research is typically fruitful. However, they also suggest that researchers remain flexible and be willing to alter their measures based on pilot data or initial attempts at coding that imply new issues not previously addressed. In essence, one should plan ahead but be open to some improvisation as the need arises. They suggest that a more focused approach to the aspects of process that are most theoretically interesting will generally lead to better results. However, they also suggest that a wide data net be cast (i.e., collect as much information about the group process as one can) so that information thought less important early on can still be assessed if later it appears more relevant. With digital recording and computer technology, keeping a complete record of all verbal and nonverbal behavior during group interaction makes following the wide-net suggestion far easier than it used to be. Finally, they suggest researchers build their coding schemes from well-formulated theory so as to ensure a degree of coherence in the analysis process. But, they also tell researchers to pay attention to their data so that interesting patterns that may not have been predicted are not overlooked.

Surveys and Interviews

Although survey and interview studies of groups are not common in social psychology, they are quite useful when appropriate, such as when the behaviors of interest can be safely assumed not to be highly dependent on the setting where responses are sought. For example, in surveys of political factions, it can usually be assumed that within fairly broad limits, the respondents' preferred policy will

not depend on the survey type (telephone, mail, in-person) or the particular setting where the faction or its representative is contacted. Another reason to rely on such methods is because it may be impractical, unethical, or even illegal to observe or manipulate the group of interest but possible to survey or interview group members afterward. For example, direct observation of actual jury deliberation is (with very few exceptions) illegal in the United States, and hence most data from such groups must rely on post-trial juror interviews.

Doing surveys or interviews of group members, for the most part, raises the same methodological concerns that arise in any survey or interview (e.g., obtaining large and representative samples, establishing rapport and avoiding respondent response biases, composing unambiguous and nondirective questions; see Bartholomew, Henderson, & Marcia, 2000; Cannell & Kahn, 1968; Hyman, 1978; Visser, Krosnick, Lavrakas & Kim, Chapter 16 in this volume)). A couple of distinctive issues that arise when group behavior is of interest are (1) how many group members must be surveyed/interviewed and (2) should group members be surveyed/interviewed separately or together. For the first question, the ideal, of course, is for every group member to be questioned, but this is often not possible for a variety of reasons (e.g., locating group members, refusal to participate). When the information sought is available to all group members and there are unlikely to be distorting response biases, only the reliability of measurement is likely to be compromised by relying on the responses of a subset of the full group. However, when only certain group members are likely to possess the sought-for information, when there are good reasons to suspect response distortions (e.g., hindsight bias, social desirability biases), or there is considerable within-group variability among members around the collective, group's response, partial sampling of the group can introduce both systematic and random error. For example, Kerr and Huang (1986) showed that a variable that accounted for a single group member's preference to some degree would account for far less (typically more than 20 times less) variance in the group's preference. This was true for a wide range of group sizes, strength of prediction at the individual level, and group decision-making processes. As to the second question, generally speaking it is preferable to survey/interview group members separately (to minimize statistical dependence and mutual social influence on responses). However, where the accuracy of memory of some event occurring in the group is paramount, the demonstrated ability of group members to catch and correct one another's memory mistakes (Betts & Hinsz, 2010) could justify questioning group members together. (Focus groups – another type of collaborative interviewing technique – is discussed in more detail later in the

chapter.)

Studies attempting to estimate the operative social decision scheme linking predeliberation juror preferences with the final verdict of actual juries can illustrate the use of survey methods to study group processes. For example, Sandys and Dillehay (1995) did telephone surveys of ex-jurors to assess the vote split at the first jury ballot. Using this method, they replicated in actual juries several results found in jury simulation experiments (e.g., that initial majorities nearly always prevail; that juries with even splits were most likely to hang; Stasser, Kerr, & Davis, 1989). Surprisingly, even on so public an event as the first ballot of the jury, there was considerable disagreement among surveyed jurors; for a sample of 50 focal trials for each of which 3 jurors' responses were sought, in only 22% of the trials did the polled jurors agree unanimously on the first ballot split. Hence, the results for a much larger sample of 190 non-focal trials (with only a single juror interviewed) were probably far less reliable.

Computer Simulations

Computer simulations are a particularly useful technique for studying groups or collective behavior more generally (Davis & Kerr, 1986). Using basic assumptions drawn from data on a variety of groups in conjunction with formal models of group processes can provide insights into how such groups might operate and how various procedural variations might influence their final judgments. A number of group research domains have put computer simulations to good use. Computer simulations of jury decision making have been used extensively to assess the potential impact of various procedural variations on jury performance (Davis & Kerr, 1986; Filkins, Smith, & Tindale, 1998; Kerr, MacCoun, & Kramer, 1996; Tindale & Nagao, 1986; Tindale & Vollrath, 1992). Using extensive data from mock jury studies to set parameters, procedural factors such as jury size, assigned decision rule, jury selection procedures, and jury instructions were evaluated in terms of their potential effects on jury verdicts. Research on social dilemmas has used computer simulations to address such questions as how cooperation can evolve in groups when defection is more individually rational (e.g., Takagi, 1999; Watanabe & Yamagishi, 1999). Recent work using evolutionary game theory approaches have shown that majority processes are very accurate (i.e., tend toward optimal choices) and extremely efficient for resolving group member preference differences (Kameda, Takezawa, & Hastie, 2003). They have also shown that in-group favoritism and out-group distrust in combination is more stable in a dynamic intergroup

environment than other possible combinations (Choi & Bowles, 2007). Computer models have also been used to study issues of diversity in small groups (Larson, 2007). Recently, multi-agent computational models have been used to simulate both transactive memory systems (Ren, Carley, & Argote, 2006) and how person perception processes influence and are influenced by individual, dyadic, and social network information, helping understand how socially shared cognitions are created and used (Smith & Conrey, 2007). Each of these examples helps both demonstrate and capture the complexity inherent in group behavior, and future work along these lines will continue to inform and enhance our ability to understand complex group interactions.

Methods for Analyzing the Structural Properties of Groups

As the preceding discussion of group observational methods suggests, a central question in the study of groups is how groups are structured – that is, what is the pattern of relationships (power, influence, status, liking, etc.) among the members of the group? A number of special techniques for analyzing group structure have been developed to address this central question.

Sociometry

A traditional method of exploring the structure property of relations among group members is Moreno's (1953) sociometric technique. It begins with each group member choosing some number of other group members preferred on one or more dimensions. The simplest (and probably most common) choice is for each group member to choose the single other group member he or she likes best, but the dimension(s) of judgment could reflect any interest of the investigator (e.g., Who are preferred coworkers? Who is most respected?). These preferences are recorded in a *sociomatrix*, where rows represent judges, columns represent targets, and the entries are the (presence or absence of) expressed preferences. Column totals summarize each target's social acceptance or *sociometric status*. Other summary indices can be derived from this matrix, such as the number of group members choosing one (social receptiveness or choice status) or the number of mutual choices in the group (as an index of group cohesiveness ; Northway, 1967).

A *sociogram*, a graphical summary of the information contained in the sociomatrix, can also be created. Every group member is designated by a

geometric shape (typically a circle, although one can represent subtypes of interest [e.g., men and women] with different shapes). Then group members' preferences (typically their first or strongest preferences on a single dimension) are indicated by arrows connecting judge to preferred target. A more easily comprehended picture of the group's structure can usually be created by rearranging the group members on the page to highlight patterns of choice (e.g., by putting a person chosen by many group members in the middle of a cluster; by putting those rarely chosen at the edges of the figure). Group members who are distinctive can be easily identified in the final sociogram. These designated individuals include those who are preferred by many group members (*stars*), those preferred by few or no group members (*isolates*), those who comprise subsets or cliques within the group that are mutually connected (*chains*), and pairs of group members that choose one another (*reciprocated pairs* or *friends*). There are also more complex statistical techniques (Cillessen, 2009; Kafer, 1976; Lindzey & Borgatta, 1954; Sherwin, 1975) and software (e.g., Levin, 1976; Noma & Smith, 1978; SociometryPro, <http://www.ledisgroup.com/en/topsocioen>) that can be used when one's data set is large or varied (e.g., containing preferences on several dimensions).

Social Network Analysis

Social network analysis is similar to Moreno's (1953) sociometric approach but is a far more flexible, powerful, and widely used method (primarily in sociology, political science, and anthropology, but in social psychology as well; cf. Katz, Lazer, Arrow, & Contractor, 2005) for analyzing a group's structural properties. Like sociometry, network analysis utilizes dyadic relationships as the basic unit of analysis, matrix summaries of the raw data, indices summarizing aspects of group members' position in the group, and occasionally (particularly for smaller groups) graphical summaries of the structure relationships. However, social network analysis has a much more fully developed set of analytic techniques (exploiting advances in graph theory) and can be applied to a much larger variety of relationships, to relationships varying in strength as well as existence, to summarizing aspects of the full network, and to structural patterns in much larger and more complex social aggregates (e.g., at the organizational, national, or international levels).

It is well beyond the scope of this chapter to provide a full overview of the techniques of social network analysis. Rather, we shall simply note a few basics of these techniques. There are a number of good introductory texts available

(e.g., Knoke & Yang, 2008; Scott, 2000; Wasserman & Faust, 1994) that interested readers can use to pursue the study of this sophisticated technique.

Network analysis begins with a set of *nodes* or *actors*. In small group research, this is likely to be the set of group members, but it could also be other objects, either social (e.g., organizations, clubs) or nonsocial (e.g., events, locations). The set of actors examined may represent a tractable and well-delimited collective (e.g., an intact group), but could also be a random or snowballed sample from some very large or amorphous collective. The basic relational data reflect the existence, nonexistence, and/or strength and frequency of relationships (or *links* or *ties*) between these actors. What kind of relationship is assessed will depend on the investigators' objectives and hypotheses, but could, in principle, be of any sort. Commonly studied relationships include sentiment (e.g., liking) relationships, exchanges of information or commodities, social influence relations, workflows, or kinship relations.

Network data can be obtained in any of several ways (e.g., from archives, by direct observation of group interaction, by self-report via questionnaire or interview). The raw data can be tabulated in any of several equivalent matrix forms. Probably the most straightforward means of compilation is the $N \times N$ (where N is the number of actors) sociomatrix described earlier. When the relational data are unidirectional or nondirectional, the matrix is symmetric, and the $N(N - 1)/2$ elements below the diagonal suffice; when the relational data are directed (i.e., Actor A's relationship to Actor B cannot be assumed to be equivalent to Actor B's relationship to A), the matrix need not be symmetric, and entries both above and below the diagonal must be specified.

Network analysis presumes that "the structure of relations among actors and the location of individual actors in the network have important behavioral, perceptual, and attitudinal consequences, both for the individual units and for the system as a whole" (Knoke & Kuklinski, 1982, p. 13). Thus, this technique seeks to relate behavior of interest to features of the network. The latter can be statistics associated with specific actors, such as an actor's number of direct links with other actors (*degree*), the ease of an actor reaching all others (*closeness*), an actor's *centrality* in the network, or relative level of being the object rather than the source of relations (*prestige*). Actors who occupy distinctive positions in the network may be assigned distinctive roles. Some of these (e.g., star, isolate) are similar to sociometric roles mentioned previously; other roles of note include an actor who connects clusters of which he or she is not a member (*liaison*), an actor who belongs to two or more clusters (*bridge*), or an actor who connects

one part of the network with another (*gatekeeper*). Other features describe a particular or the average link, such as its temporal stability, symmetry, or directness. Such analyses can be extended to focus on aspects of a particular or the average triad (e.g., What's the degree of transitivity of links?). Finally, the analysis may focus on features of the entire network, such as its size, the average path distance between actors (*connectivity*), the ratio of mutually reachable pairs of actors to all possible pairs (*connectedness*), the relative centrality of the most central actor to all other actors (*centralization*), the ratio of connected to possible links (*density*), and so forth. Such analyses are aided by social network analyses software packages (see Hansen, Shneiderman, & Smith, 2010; Scott, 2000; http://en.wikipedia.org/wiki/Social_network_analysis_software). Study of the range of applications of social network analysis (e.g., Scott & Carrington, 2011) can provide a fuller appreciation of this technique's power and versatility.

Innovative Methods and Tools for Group Research

Traditionally, research on small group processes has been a fairly low-tech affair. For example, early observation of group process (e.g., Stephan & Mishler, 1952) relied on online coding by live observers. Clearly, both the quantity and quality of data that could be obtained were severely limited. Similarly, manipulation of interesting features of groups' environment, structure, or process was generally crude and intrusive in many early studies. For example, the structure of group communication might be varied by physically arranging group members so that written notes could be passed physically only through certain slots (e.g., Guetzkow, 1968). The apparent content of intermember communications might be manipulated by the investigator originating or intercepting and replacing such written notes (e.g., Schachter, Ellertson, McBride, & Gregory, 1951).

The rapid growth of technology during the last few decades has certainly increased the potential for more detailed, reliable, varied, and sophisticated small group research. In what follows we describe a number of the particular ways in which modern technology has been and could be applied to such research. We make no claims that this overview is comprehensive, which is precluded, in part, by the fact that new types of hardware and software are appearing regularly; "cutting edge" technologies can become obsolete in even the relatively short time lag between writing a chapter and its publication. We also wish to stress that whenever we mention a particular piece of technology, we do so only to illustrate how technology has been or might be applied, and not as an

endorsement. Interested readers should take any of our illustrations only as starting points, and undertake their own investigation into the advisability of applying any particular technology to their own particular substantive questions. To aid in such investigations, we occasionally provide Internet links that contain and maintain product descriptions, reviews, and other sources of relevant information. (Also note that although these websites appear useful at present, they may or may not continue to be in the future.)

Audio-Video Hardware and Software

As we mentioned earlier, arguably the most important technological innovations for observational research on small groups is the development of reliable, affordable, compact, and easy-to-use equipment to make audio or videorecordings of group interaction. Of course, audiotaping or filming group interaction has technically been possible since the advent of modern social psychology, but these technologies either lost much information that was of interest (e.g., identity of speaker, target of communications, all other overt nonverbal behaviors in the case of audio recordings) or were expensive and cumbersome to use (in the case of film and early, reel-to-reel video). However, with the advent of compact video cameras and digital recording, it has become fairly simple and inexpensive to make high resolution videorecordings of group behavior.

Earlier we noted some of the advantages of videorecording over live observation – for example, multiple observers and investigators can examine and code the same interactions at their convenience and with less risk of fatigue, slow-motion replay can reveal subtle or easily missed behaviors, and distracting or biasing information can be masked. Easily available video technologies such as remote camera controls, video-mixing boards, and video-editing hardware also make it feasible to focus on particular and subtle aspects or combinations of observable behavior (e.g., a particular group member, simultaneous actions of a speaker and listener).

With or without permanent videorecordings, observational research of group behavior can be labor intensive. However, there are also a number of technologies currently available that make the task less onerous and more flexible. For example, several computer programs (e.g., The Observer XT) enable one to use the computer keyboard to encode multiple events of interest in real time. These are particularly useful where videorecording is not feasible for reasons of practicality (lack of hardware) or methodology (e.g., the use of a

camera would be intrusive or unethical). There are also a number of hardware/software packages (e.g., MacSHAPA, Anvil, ODCS, CowLog, The Observer XT; Hänninen & Pastell, 2009; MacLin & MacLin, 2005; Noldus et al., 2000; Sanderson, 1994; Tapp & Walden, 1993; <http://academic.csuohio.edu/kneuendorf/content/cpuca/avap.htm> or http://bama.ua.edu/~wevans/content/csoftware/software_menu.html) that are designed for coding data from videotape or digital video files. Such software can tally not only particular events but other interesting features (e.g., durations) as well. Some of this software also permits the integration and synchronization of multimodal signals from various sources, such as observational, video, tracking, and physiological data (Zimmerman, Bolhuis, Willemsen, Meyer, & Noldus, 2009). When the research is at an exploratory stage, several computer-assisted qualitative data analysis software (CAQDAS) packages (e.g., see http://en.wikipedia.org/wiki/Computer_assisted_qualitative_data_analysis_software) are also available.

Such programs can include a number of useful features, such as large numbers of possible coding categories, keyboard control of the video source, precise timing of event occurrence and duration, visual or auditory feedback of entered codes, and the ability to annotate event coding. Thus, rather than coding a single variable through laborious procedures (e.g., manually rewinding, using a recorded timer or visual content to find the start of the event), by using such technology one can simultaneously code several features of interaction, mark and automatically return to points of interest, and use feedback features to detect unanticipated patterns in the data. One can also either do a number of standard (e.g., interjudge reliability) or not-so-standard (e.g., lag sequential analyses, transition analyses, analyses of cyclic activity; see Bakeman, 2000; Heyman, Lorber, Eddy, & West, Chapter 14 in this volume) analyses within such programs or export the data for analysis with other statistical packages.

Such video software still requires the decisions of human judges. For certain simple aspects of group interaction, one may design equipment to obviate the human judge. For example, Dabbs and Swiedler (1983) developed a system for automatically monitoring the onset and ending of speech in group discussions. As technological advances occur in shape, movement, and voice recognition by computer, it is likely that it will be possible to automate many other coding tasks (e.g., see Cohn & Sayette, 2010 for coding facial expressions), which should bring attendant gains in accuracy and efficiency of coding.

Computer Technology: Data Collection at Arbitrary Group Tasks

An even more revolutionary technological innovation of the late 20th century for social psychology (as for nearly every other discipline, as well as for the general public) is certainly the development of powerful, small, and affordable microcomputers. Here we focus our attention on how the computer can and might be used as a tool for conducting group research.

Three generic approaches to computer-mediated experimentation on groups might be distinguished for our immediate purposes.

1. The first approach has a group working together at a single computer. In this setting, the computer serves as an instruction and/or stimulus-presentation device, and/or as a data-recording device (typically for group responses through the keyboard, but possibly for individual member responses [e.g., by taking turns] and via other input devices, such as joysticks, analog/digital boards, etc.). For example, rather than have a single pad for recording ideas generated by a brainstorming group (cf. Diehl & Stroebe, 1987), one could provide the group with a computer to record ideas, making possible richer data collection (e.g., the rate as well as the number of ideas generated).
2. The second approach has each member of a real or purported group working at separate, stand-alone computer stations. This approach is particularly appropriate for research questions about those group processes that do not involve any actual interpersonal activity (e.g., social facilitation) or that, at most, involve restricted patterns of interaction (e.g., a context in which group members are allowed to talk to one another as they work at their computers; cf. Olson, Olson, Storreston, & Carter, 1994), but it can also be used for certain group simulations where the experimenter programs in and controls the apparent responses or communications of other group members. For example, Messick, Wilke, Brewer, Kramer, Zemke, & Lui (1983) led participants to believe that they could monitor each other's harvests from a shared resource pool through computer feedback. In fact, there was no feedback of actual choices, but rather false feedback preprogrammed by the experimenters to examine participants' reactions to various patterns of resource use (e.g., a steadily declining resource pool, high vs. low variance in members' harvests).

3. The third approach provides each group member with his or her own station and permits intermember communication via a computer network. A striking example of this approach is Latané and L'Herrou's (1996) study of different allowable communication links – modeling different spatial arrangements of group members – and their effect on patterns of social influence. The use of asynchronous computer communication (e.g., e-mail) allowed these investigators to both control channels of communication and overcome the difficult logistic problem of composing 24-person groups for several rounds of communication.

There are, in turn, several generic means of acquiring the software needed to undertake these approaches:

1. One can identify and obtain existing software. There are many such application-specific programs that have been developed for small group research (e.g., in social dilemma research, see Messick et al., 1983 for an illustration). Such software is usually identified through careful study of the existing empirical literature, by word of mouth, or by examining databases of psychological software (cf. <http://www.psychology.org/links/Resources/Software/>; <http://psych.hanover.edu/Krantz/software.html>; <http://www.psywww.com/resource/bytopic/software.html>). Of course, the chief drawback of using preexisting software is that it is generally inflexible, not permitting alterations in procedure or experimental parameters. In a few cases, investigators have tried to build flexibility into their programs so that other investigators could adapt them to new purposes. Illustrations are CDS (Li, Seu, Evens, Michael, & Rovick, 1992), which captures typed dyadic communication, and GROUPCOM (Levine, 1978), which permits interpersonal communication among up to six group members. Another, related option is to use widely available chat rooms or instant messaging services (e.g., AIM, Google Talk, Skype) to structure asynchronous or synchronous group interaction. Such services either have their own options for recording text, audio, or video content, or one can obtain add-on software (e.g., Hotrecorder, MX Skype) and hardware (e.g., a video capture card) to record such content for later analysis.
2. If one is (or can afford to hire) a talented computer programmer, one can program one's computer or computer network and apply any one of these approaches to one's substantive research question. This approach,

of course, carries maximal flexibility, but is beyond the training or resources of many investigators.

3. There also exist a number of general-purpose programs developed specifically for psychological experimentation. Several such packages were developed early on by experimental and cognitive psychologists for the Mac platform (e.g., Chute, 1993; Cohen, MacWhinney, Flatt, & Provost, 1993; Haxby, Parasuraman, LaLonde, & Abboud, 1993; Hunt, 1994; Vaughan & Yee, 1994). Today, there are several such programs available for the Mac (e.g., PsyScope, SuperLab) or the PC platform (e.g., MediaLab, DirectRT, E-Prime, SuperLab, Inquisit, Authorware). These packages typically include many useful tools for conducting experiments, such as options that permit counterbalancing orders of stimulus presentation, precise timing of stimulus and response, and so on. Unfortunately, at present, none of them is designed to take advantage of computer networking, so that they can typically only be used for those applications without actual interaction among group members. Although, to our knowledge, there currently is no general-purpose experiment generator that is networked, there have been attempts to extend general-purpose authoring software from use for stand-alone experimental applications (e.g., Wolfe, 1992) to networked applications (e.g., Hoffman & MacDonald, 1993). Currently, one can incorporate Web-based applications (e.g., chat rooms) as stand-alone segments within a MediaLab questionnaire/experiment. There is also a theoretical capability of capturing the data collected in such applications and integrating them with those collected directly by MediaLab, although this capability is not yet well developed (Jarvis, 2011).
4. The market for sales of hardware and software for all of experimental psychology is, compared to the larger IT market, relatively small (Schneider, 1991). Consequently, little research and development in the computer industry has focused on the requirements of psychological researchers in general, let alone those interested in the study of small group behavior in particular. However, there is both a considerable market for and commercial interest in technology that aids in interpersonal communication – what McGrath and Hollingshead (1994) generically termed group communication support systems (GCSSs) – and that assists organizational groups or teams to improve their productivity – group performance (or decision) support systems (GPSSs; McGrath & Hollingshead, 1994). So, a final means of applying technology to the study of group process is to directly utilize or adapt

technology developed for these more applied purposes as group research tools.

GCSSs are simply tools for extending human communication beyond its most basic form (viz., face-to-face verbal/nonverbal interaction). GCSS technologies currently exist that permit synchronous or distributed (in both time and space) communication via various modalities (audio, video, video and audio, typed text, handwritten text, graphics; McGrath & Hollingshead, 1994). These technologies range from the mundane (telephones, surface mail), to the commonplace (e.g., cellular phones, voice mail, electronic mail), to the relatively novel (e.g., interactive chat rooms, video conferencing via the Internet; see <http://thinkofit.com/webconf/> or <http://www11.informatik.tumuenchen.de/cscw/> for introductions to a few of the possibilities currently available). Although such GCSSs are not commonly used as tools in small group research at present, we believe that they have considerable potential to be used in this way (see McGuire, Kiesler, & Siegel, 1987 or Hollingshead, McGrath, & O'Connor, 1993, for illustrations of this potential). In organizational settings, this potential is already being realized in the burgeoning literature on *virtual teams*, whose members may be geographically dispersed as they undertake their collective tasks. Such virtual teams have provided a new and fascinating context wherein research questions about group dynamics and performance can be posed and answered (e.g., Curseu & Wessel, 2008; Hertel, Geister, & Konradt, 2005).

Although social psychologists have not put GDSS or other technological innovations to much use as research tools, the recent interest in teams in engineering and technology has led to a few interesting examples (e.g., Matsatsinis, Grigoroudis, & Samaras, 2005; Paul, Haseman, & Ramamurthy, 2004). More collaborative work between social psychology, engineering, and information technology researchers would probably lead to new and interesting ways for GDSS systems to be used as research tools.

GPSSs attempt to do more than simply facilitate communication among group members. They attempt to restructure common group tasks, often incorporating innovative communication technology, so as to enhance group productivity. GPSS technologies have given birth not only to an industry aiming to exploit the commercial possibilities of such systems (e.g., <http://www.ventana.com>) but also to a burgeoning group of scholars with sophisticated research centers (<http://www.uasabilityfirst.com/groupware>), major conferences (e.g., the biennial Computer Supported Cooperative Work [CSCW] meetings; the annual

Human Computer Interaction [HCI] meetings; cf. <http://www.acm.org/events/>), and specialized scientific journals (e.g., *Communications of the ACM*, *Information Systems Research*).

One product of this marriage of commercial and scholarly pursuits is a rich empirical literature (McGrath & Hollingshead, 1994). Another is an impressive and varied collection of “groupware” – hardware and software products designed to facilitate collaborative work (see <http://www.telekooperation.de/cscw/>) – ranging from collaborative editing tools to message systems to group meeting support systems to conferencing systems. Ventana Corporation's GroupSystems package is an illustration of a GPSS. It contains modules for generating and categorizing ideas, outlining topics, commenting on ideas, and evaluating and voting on proposals.

Of course, the scientific study of group performance has been a major topic of social psychology since its inception (Kravitz & Martin, 1986). It is thus a bit surprising to find so few social psychologists actively involved in the study and application of technology to group work (see Kielser, 1997 and McGrath & Hollingshead, 1994 for noteworthy exceptions). We suspect that these emerging disciplines hold tremendous potential not only to provide us with useful tools for controlling and observing group behavior but also to raise fascinating new questions about group behavior, which would never occur to us without the many new possibilities for structuring group work that modern technologies create. The study of brainstorming in electronically linked groups, described in a later section, is an excellent illustration.

Groups as a Context/Mean for Research and Application

Thus far we have been emphasizing methodological tools that are useful when the primary goal is the study of group behavior per se. In this section, however, we shift focus somewhat. Here we examine a number of methodologies in which some guided form of group interaction has been held to provide a useful context and means for achieving some other goal, such as solving a problem, assessing opinion, generating ideas, and so on. In effect, these are also “group productivity/decision support systems,” but ones that usually require no exotic technologies. For the most part, these methodologies have not been developed, nor are they commonly used, by social psychologists; in these senses, they represent innovative group techniques. And, for the most part, there is little

conclusive research evidence on the efficacy of these techniques. However, because they are employed (at times quite widely) outside social psychology and because the use and goals of these techniques pose a number of interesting and patently social psychological questions, we have chosen to describe them here. We have been somewhat selective, however; in particular, we have excluded methods of using groups for various therapeutic ends (see Forsyth, 2009, chapter 16, for an introduction to the latter methods).

In what follows we briefly present the genesis, rationale, basic procedures, a sourcebook or two, and (when available) evidence for efficacy for each of the following: group brainstorming, focus groups, quality circles, nominal group technique, the Delphi method, and judge-advisor systems. These are roughly ordered in terms of increasing structure and constraint on interpersonal interaction.

Group Brainstorming

Brainstorming was developed by advertising executive A. F. Osborn (1957) as a means of facilitating the generation of creative ideas through face-to-face group interaction. Osborn prescribed four rules for such brainstorming groups. First, members are instructed to express any ideas that come to mind without concern for their quality, practicality, and so forth. Spontaneous and uninhibited “free-wheeling” is encouraged. Second, during brainstorming there should be no evaluation of any ideas expressed. Emphasis should be entirely on the generation of ideas, not their evaluation. Third, the brainstorming group should strive for as many ideas as possible; the more ideas, the better. Fourth, group members should try to build on others’ ideas, combining, improving, and extending wherever possible.

Osborn (1957) made rather extravagant claims for the efficacy of group brainstorming – for example, “the average person can think up twice as many ideas when working with a group than when working alone” (Osborn, 1957, p. 229). Unfortunately, systematic research has failed to substantiate these claims. To the contrary, a sizeable literature (see Diehl & Stroebe, 1987, Mullen, Johnson, & Salas, 1991, and Nijstad, 2009 for reviews) has consistently shown that brainstorming groups produce both fewer and poorer-quality ideas than equal-sized, identically instructed nominal groups (i.e., groups whose members work in isolation and whose total output is determined by pooling members’ output, eliminating any redundant ideas).

Substantial progress has been made in identifying the sources of this process

loss in brainstorming groups, with production blocking (i.e., the fact that only one person can talk [and, perhaps, think] at a time in the face-to-face group), production matching (i.e., social comparison and modeling of low levels of productivity), and evaluation apprehension (i.e., fear of negative evaluation for voicing ideas in the group context) all emerging as contributing processes (Diehl & Stroebe, 1987; Paulus & Dzindolet, 1993; Stroebe & Diehl, 1994). Hence, procedural variations that neutralize these mechanisms (e.g., individual recording of ideas, including periods of silence, taking turns) may close the gap between nominal and brainstorming groups (Philipsen, Mulac, & Dietrich, 1979; Ruback, Dabbs, & Hopper, 1984). Also, the standard brainstorming rules can be better realized by training a group facilitator to minimize production blocking and evaluation apprehension; such a facilitator can reduce or even eliminate the usual process loss (Offner, Kramer, & Winter, 1996; Oxley, Dzindolet, & Paulus, 1996).

The most exciting procedural innovation in brainstorming is so-called electronic brainstorming (EBS). Each group member has a terminal that is networked with all other terminals. Group members type in ideas at will. At any time, a group member can see a sample of the ideas generated by the group simply by hitting a key; by repeatedly doing so, he or she can examine all the ideas generated so far. Because ideas are not attributed to particular group members, member anonymity is maintained. Recent research suggests that for small-to moderate-sized groups (less than 10 persons), EBS groups perform as well as comparably sized nominal groups, and for larger groups (around a 12 or more), the EBS groups actually outperform the nominal group baseline (Dennis & Valacich, 1993; Dennis & Williams, 2003; Valacich, Dennis, & Connolly, 1994). Such apparent “process gain” – group performance exceeding the group's apparent potential productivity – has been very rare in the social psychological literature (e.g., Laughlin, Hatch, Silver, & Boh, 2006) and is of special interest for theory development and application. The source of this apparent process gain in EBS groups appears to arise from the stimulating effect of exposure to others' ideas (Leggett-Dugosh, Paulus, Roland, & Yang, 2000; Nijstad, Stroebe, & Lodewijkx, 2002) and to the benefits of heterogeneity/diversity in idea-generating groups (Stroebe & Diehl, 1994).

Focus Groups

The *focus group* has been used most in marketing and advertising research. It is a qualitative, semistructured interview technique in which a small group,

typically 8–10 people, discusses a topic of interest under the supervision of a moderator. The information sought is usually fairly narrowly delimited (e.g., How do consumers react to a new product or product idea? How is a product actually used? How do competing products compare?). The information gleaned from focus groups may directly guide decision making or may prompt more systematic and quantitative techniques.

Considerable preparation should precede focus group sessions. The objectives of the sessions first need to be specified – what information is desired? The moderator(s) must be selected and briefed on the objectives. A moderator guide must be prepared. This guide is a detailed outline of topics that should be covered in the focus group, when each might be addressed, and how available time will be used. The appropriate respondent population must be identified and a method of participant recruiting chosen. Because the sample sizes of focus group studies, even those including several groups, are rarely large, and quantitative data (e.g., population estimates with confidence intervals) are not sought, a probability sample of the target population is usually not attempted. Consequently, generalization to larger populations is problematic. Instead, certain participant characteristics are specified (e.g., women between 30 and 45 years of age who regularly use a particular product) and the groups are then composed of samples of paid volunteers obtained in any of several ways (e.g., from community groups, via telephone or mail screenings, from firms providing names). For a number of reasons (to avoid distractions, to target specific respondent populations), focus groups are typically fairly homogeneous demographically. If information is sought from diverse subpopulations (e.g., men and women, old and young), this is typically achieved by running separate homogeneous focus groups.

Focus group sessions follow no specific set of procedures. However, in practice, there are a number of common features. Although they sometimes are conducted via teleconference, the discussion is nearly always conducted face to face and is recorded; these days, videorecordings are the norm. The moderator leads the focus group through usual stages of group discussion: general orientation (introductions, ground rules), orientation to the topic (via more general discussion), focus on specific topics of interest (defined in the moderator guide), and wrap-up. The moderator attempts to act as a facilitator, encouraging and guiding but not dominating discussion. Any of a number of mechanical (e.g., presenting product samples or commercials; having respondents write down ideas before discussion) and social (e.g., soliciting views of quiet participants; seeking reactions to most active participants) methods can be used in this

pursuit. Several special steps may be taken with unusual respondent groups (e.g., children, experts). There may be post-group discussions among investigators (e.g., the moderator and the client). There may also be a formal report prepared by the moderator to summarize and interpret the content of the focus group discussion. There are also several variants of the generic focus group, including two-way groups (where two groups may observe and comment on one another's interaction), dual-moderator groups, dueling-moderator groups (where a pair of moderators take opposing positions), client-participant groups (where one or more client representatives participate), and virtual focus groups (using telephonic or video links).

The purported benefits of the focus group technique include the following: (1) it can often be easier and less expensive to use focus groups than more traditional survey or interview techniques (although the cost per respondent can be considerably higher for some focus groups); (2) the group setting can provide insights into social forces of interest (e.g., peer pressure on product use); (3) the group setting permits reactions not only to questions from the moderator but to the comments of other group members; (4) the group setting encourages greater honesty, spontaneity, involvement, and thoroughness of responding; and, consequently, (5) one has access to more useful information, including respondents' emotional reactions, vivid anecdotes, novel ideas, vernacular expression, and so forth. Unfortunately, such claims, as well as prescriptions for focus group practice, are based primarily on *experienced validity* – the subjective evaluations of focus group users and proponents. There is very little published research documenting these claims (e.g., Bristol & Fern, 1996, 2003; Seal, Bogart, & Ehrhardt, 1998). Moreover, there clearly are limits to the applicability of focus groups – for example, when the topics are considered private and anonymity is desired, or when one wants to generalize to broad populations. However, if the purported benefits could be verified, focus groups might provide an effective technique for a variety of objectives: assessing attitudes, probing for suspicion postexperimentally, doing introspective process analyses of social processes, or for exploratory hypothesis-generating research (see Fern, 2001, Krueger & Casey, 2008, Liamputtong, 2011, or Stewart, Shamdasani, & Rook, 2006 for more detailed descriptions of focus group methods).

Quality Circles

Quality circles (or quality control circles – QCs) are used primarily in business and industrial settings. They are seen as an alternative to more traditional and

hierarchical systems of management, an alternative that involves workers themselves more actively and directly in their work and organization. QCs were developed in the 1960s in Japan and have grown in popularity in many Western industries.

Hutchins (1985, p. 1) defined a QC as

[A] small group of between three and twelve people who do the same or similar work, voluntarily meeting together regularly for about one hour per week in paid time, usually under the leadership of their own supervisor, and trained to identify, analyze, and solve some of the problems in their work, presenting solutions to management and, where possible, implementing solutions themselves.

To this end, a number of group techniques and principles are incorporated into QC procedures. For example, heavy reliance is placed on group brainstorming techniques for identifying workplace problems and solutions, the groups are limited in size to permit general participation in face-to-face meetings, and decision making is democratic – one person, one vote. Various aspects of the QCs' functioning are not distinctively social in nature, such as collecting relevant data, analyzing the causes of workplace problems, and preparing clear and persuasive presentations of recommendations to management. Implementing QCs and achieving their purported benefits (described later) are not simply a matter of forming groups of coworkers, but require fairly extensive organizational commitment and support (e.g., a willingness to invest organizational resources, a willingness to seriously consider QC proposals).

The participation and involvement of workers achieved through QCs is alleged to have extensive benefits: reduced turnover, fewer grievances, improvements in productivity, improvements of quality, higher worker morale, and stronger corporate loyalty and identification. Attempts to verify these claims empirically have produced mixed, negative, or null results (Barrick & Alexander, 1987; Park, 1991; Pereira & Osburn, 2007; Steel & Shane, 1986), and there are indications that the effectiveness of QCs is strongly moderated by other factors (e.g., the duration of the QCs; management's attitude toward QCs; see Park & Golembiewski, 1991). Although there are very difficult methodological problems in the evaluation of QCs (e.g., participant self-selection, reliance on quasi-experimental designs), the growing popularity of QCs and several indications of positive results certainly justify more careful empirical attention. Besides posing interesting substantive questions for research

on group and organizational processes, QCs might be usefully applied within research teams themselves. (See Hutchins, 1985 or Ingle, 1982 for more detailed descriptions of QCs.)

Nominal Group Technique

The nominal group technique (NGT), developed by Delbecq and Van de Ven (e.g., Delbecq, Van de Ven, & Gustafson, 1975), was designed to overcome certain aspects of unconstrained face-to-face discussion that can interfere with effective group problem solving and decision making. Of particular concern were those small group processes that tend to prevent full and thorough participation by all group members. These included (1) the reluctance of some members to participate, especially in larger groups; (2) domination of group discussion by an opinionated, loquacious, repetitive, or high-status individual or faction; (3) the diversion of time and effort to organize and maintain the group that might be devoted to generating and evaluating ideas; (4) getting stuck on a single line of argument for long periods of time; and (5) hurrying to reach a speedy decision before all relevant information has been considered. NGT attempts to counter such problems by using nominal groups (as described earlier for brainstorming research) for idea generation.

Another set of problems can arise from explicit requirements or implicit pressures to achieve consensus in groups. Group members might (6) become overcommitted to their initial publicly expressed opinion (cf. Kerr & MacCoun, 1985), (7) decline to participate or defend a position to avoid social sanctions from a leader or the majority faction, or (8) compromise or shift position simply to avoid such sanctions. NGT attempts to minimize such problems by having no explicit consensus requirement or decision rule and by pooling preferences statistically to define a group product.

Formally, there are four stages in the NGT. First, a moderator poses the problem to a group. The members of the group are given time (typically 10–20 minutes) to silently write down as many ideas or solutions as they can, much as the nominal groups used in brainstorming research. It is recommended that the group be large enough to generate a substantial pool of ideas but not too large to make the following stages unwieldy; 7–10 members are thought to be optimal. During the second stage, group members state the ideas that they have written using a round-robin procedure. After each idea is stated, the moderator writes it down on a blackboard or flip-chart. Stage 3 consists of open group discussion of the recorded ideas. The emphasis here is on clarifying and evaluating each idea;

there is no goal of consensus. A group decision or a preference ordering for ideas is determined by a nominal voting procedure at the fourth and final stage. Nominal voting requires each group member to privately evaluate the alternatives (e.g., rank ordering one's favorite five alternatives). The moderator pools these evaluations (e.g., computes mean rank orders) to identify the group's overall preference(s). Optional additional stages are another group discussion (this time focusing on the group decision) and another vote.

Proponents of the NGT take the sizeable literature demonstrating the superiority of nominal to brainstorming groups as indirect evidence for a superiority of the NGT to normal, face-to-face groups for idea generation. Van de Ven (1974) confirmed this claim empirically and also found that group members were more satisfied under a NGT than free interaction, a finding that he attributed to fuller, more uniform input under the NGT (cf. Stephenson, Michaelson, & Franklin, 1982). There is also some evidence that allowing group members first to share likelihood-ratio estimates before group discussion (consistent with Stages 1 and 2 of the NGT) produces more accurate aggregated post-discussion estimates (Gustafson, Shukla, Delbecq, & Walster, 1973), relative to groups without such pre-discussion sharing. It seems fair to conclude that the evidence for the NGT, although fragmentary, is encouraging (e.g., Arunachalam & Dilla, 1995; Delbecq, Van de Ven, & Gustafson, 1986; Duggan, 2003; Frankel, 1987; Henrich & Greene, 1991). The availability of GCSSs also raises new opportunities to examine innovative modifications of the traditional NGT, much as it has for group brainstorming (e.g., Dowling & St. Louis, 2000; Lago, Beruvides, Jian, Canto, Sandoval, & Taraban, 2007). (See Delbecq et al., 1986 and Korhonen, 1990 for more detailed descriptions of the nominal group technique.)

Delphi Technique

The Delphi technique seeks to pool the opinions of a group of people who are well informed or expert on some topic of interest, but without direct, face-to-face interaction. Rather, an iterated sequence of questionnaires is sent to the group by a monitor. The monitor (who could be an individual or a project team) is the conduit through which all communications are channeled. The monitor begins by identifying a panel of experts to whom an initial questionnaire is sent. In addition to dealing with several preliminary issues (e.g., explaining the project's purposes and procedures, seeking respondent commitment to the project), the initial questionnaire poses some root questions on which subsequent rounds of

the procedure are built. These questions would typically be few, very general, and open-ended; the goal is to let the group members (and not the moderator) define the domain of relevant opinions or issues. After the questionnaires are returned to the moderator, his or her next task is to develop a new questionnaire that (1) accurately and objectively summarizes group members' opinion from the initial questionnaire and (2) poses a revised, more focused set of questions for the next round. The new questionnaire is then sent back to group members. The feedback from the previous round keeps group members' identities anonymous and should ideally provide more than indices of central tendency. For example, in a Delphi application seeking technology forecasts, respondents might be given the median and interquartile range for estimates of when each of several events is expected to occur (e.g., "when will 90% of all university faculty have and use electronic mail?"), along with summaries of the supporting arguments provided by advocates of high, middling, and low estimates. Ideally, the procedure of questioning, summarizing responses, and re-questioning is repeated as long as there seems to be progress (e.g., opinion continues to converge, positions are not static). At least two rounds are required for Delphi; the original developers recommended four rounds as optimal.

The Delphi technique was developed at the Rand Corporation (Brown, 1968; Dalkey & Rourke, 1971; Helmer, 1966) as a means of pooling expert opinion. It has often been used to make technological forecasts (Rowe & Wright, 1999, 2001), but is not restricted to such tasks; "it can be used for any purpose for which a committee can be used" (Martino, 1983, p. 16). It is seen as particularly useful when informed yet subjective judgments are the only or best data available for decision making, when face-to-face discussions are impractical (e.g., because the best-informed respondents are numerous, dispersed, or hard to schedule), or when one wants to avoid certain social psychological consequences of face-to-face discussion, which are presumed to undermine effective decision making (e.g., see the factors listed in our discussion of the NGT earlier in the chapter).

The Delphi technique also has drawbacks. It requires respondents to complete and return several questionnaires. This requirement is likely to be a special problem when group members are busy (as genuinely expert respondents are likely to be) and the questionnaires seem complex or the iterated versions seem redundant. The process can also be expensive and time consuming (typically taking at least a few weeks when mail questionnaires are used); the advent of computer-mediated communication has helped reduce the latter problems. (See Kerr & Tindale, 2011 for an analysis of Delphi's strengths and weaknesses

relative to alternative group aggregation methods.)

A final problem with Delphi, as with several of the other techniques described here, is that there is little empirical research documenting its efficacy. There are some suggestive findings (e.g., Dalkey, 1968, 1969–1970; Rohrbaugh, 1979); Dalkey reported that Delphi was superior to face-to-face interaction group estimates for almanac-type questions, but the validity and generality of the claims made for Delphi await systematic research attention. (See Alder & Ziglio, 1996, Delbecq et al., 1986, Keeney, McKenna, & Hasson, 2011 for more detailed descriptions of the Delphi technique. See Sackman, 1975 for a pointed critique of the method.)

Judge Advisor Systems

A relatively recent technique for both simulating some real group decision settings and furthering our understanding of social influence processes in groups is the Judge-Advisor Systems approach (Sniezek, 1992; Sniezek & Buckley, 1995). In many settings, final decisions are made by an individual person or judge (military leader, CEO, etc.), but only after soliciting advice from a number of others (advisors). Sniezek (1992) argued that such decisions are a group product, and by conceptualizing group decisions in this way, one could attempt to isolate the influence of each person (either judge or advisor) on the decision outcome. By manipulating the amount of information advisors could provide for judges (action preference, confidence level, rationale, etc.) and the number of advisors, Sniezek and her colleagues have attempted to assess how judges used advice in making decisions.

Variation among advisors in expertise, past performance accuracy, and stated confidence can be observed or created in order to assess how each factor influences the final decision by the judge. Two relatively robust findings from this approach are that an advisor's stated confidence is a strong predictor of influence (Sniezek & Buckly, 1995), and that judges tend to weigh their own preference more heavily than that of their advisors, even when the advisors have more expertise and accuracy (Harvey, Harriea, & Fischer, 2000; Yaniv & Kleinberger, 2000). Judges are also more influenced by advisors who tend to agree with them (Harvey et al., 2000).

Afterword: On the Illusion of Group Effectiveness

A curious anomaly has been reported by brainstorming researchers (Paulus &

Dzindolet, 1993; Stroebe, Diehl, & Abakoumkin, 1992). Although interacting brainstorming groups consistently perform less well than comparable nominal groups, participants in both conditions believe that they are, and have been, more productive in a group than working alone.

In this section we have considered a number of methods, all of which extol the particular effectiveness of group settings for accomplishing varied tasks. And indeed, as Steiner (1972) has shown theoretically, for most tasks the potential productivity of a group is greater than mean individual productivity. The illusion of group effectiveness documented in brainstorming groups may stem (in part or in whole) from some confusion between what the average individual can do and what a nominal group of such individuals can do. It may also stem from there being more instances, when working alone, of feeling stumped or unsure how to proceed; the higher rate of such apparent failures can also explain the greater task enjoyment and satisfaction observed in brainstorming groups, compared to nominal groups (Nijstad et al., 2006). It is important to keep this illusion in mind when considering group methods that are highly touted but inadequately evaluated.

Conclusions

We hope that we have been able to show that the distinction between individual and group phenomena is an important one. Group processes are fundamentally different from individual psychological phenomena in important ways. In this area of social psychology (as well as related areas, such as the study of interpersonal relationships) we must examine the behavior of individuals as they are simultaneously being affected by the overt or implicit behaviors of others. The investigator must ensure that his or her methods create a truly “social” experience. Hence, to study group and other interpersonal phenomena routinely requires not only a different, more complex set of concepts and units of analyses but also a different, more complex set of methods than is needed to study individual behavior.

Allport (1962) suggested that the contrast of individual and group behavior represents the master question of social psychology. Steiner (1986) has suggested that the dominant meta-paradigm of social psychology at the end of the 20th century featured individual-level analyses and focuses on single-factor, intrapsychic, cognitive mediators of behavior. He argues persuasively that this meta-paradigm is inimical to the study of group phenomena. The many forces –

theoretical, professional, and cultural – that have produced this meta-paradigm (cf. McGrath & Altman, 1966; Steiner, 1986) are powerful and show no signs of abating. Yet as scientific social psychology enters its second century, we continue to be optimistic that it will not lose sight of the master question that dominated the initial decades of its first century. Analyses of publication trends (Moreland, Hogg, & Hains, 1994; Wittenbaum & Moreland, 2008) have suggested that interest in group phenomena has been increasing after several decades of decline. A hopeful sign is that much of this new interest reflects the integration of traditional topics of intragroup process (see Table 9.1) with some topics that have received much attention during social psychology's past few decades, such as social cognition and intergroup relations. Although the study of group phenomena does present a number of special difficulties, both conceptual and methodological, whether these trends continue will have less to do with overcoming such difficulties than with how clearly we recognize the centrality of group phenomena for human social behavior and accept the challenge of tackling the master question of our field.

References

- Alder, M., & Ziglio, E. (1996). *Gazing into the oracle: The Delphi method and its application to social policy and public health*. London: Jessica Kingsley.
- Allport, F. (1962). A structuronomic conception of behavior: Individual and collective: I. Structural theory and the master problem of social psychology. *Journal of Abnormal and Social Psychology*, 64(1), 3–30.
- Altman, I., & Haythorn, W. W. (1967). The effects of social isolation and group composition on performance. *Human Relations*, 20, 313–340.
- Anderson, A. B. (1975). Combined effects of interpersonal attraction and goal-path clarity on the cohesiveness of task oriented groups. *Journal of Personality and Social Psychology*, 31(1), 68–75.
- Anderson, B. F. (1966). *The psychology experiment: An introduction to the scientific method*. Belmont, CA: Wadsworth.
- Aronoff, J. (1967). *Psychological needs and cultural systems: A case study*. New York: Van Nostrand.
- Aronoff, J., & Messé, L. A. (1971). Motivational determinants of small group-structure. *Journal of Personality and Social Psychology*, 17, 319–324.

- Arunachalam, V., & Dilla, W. N. (1995). Judgment accuracy and outcomes in negotiation: a causal modeling analysis of decision-aiding effects. *Organizational Behavior and Human Decision Processes*, 61(3), 289–304.
- Asch, S. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.), *Groups, leadership and men* (pp. 177–190). Pittsburgh, PA: Carnegie Press.
- Back, K. (1951). Influence through social communication. *Journal of Abnormal and Social Psychology*, 46, 9–23.
- Bakeman, R. (2000). Behavioral observation and coding. In H. Reis & C. Judd (Eds.), *Research methods in social psychology: A handbook* (pp. 138–159). New York: Cambridge University Press.
- Bales, R. F. (1950). *Interaction process analysis*. Reading, MA: Addison-Wesley.
- Bales, R. F. (1965). The equilibrium problem in small groups. In T. Parsons, R. F. Bales, & E. A. Shils (Eds.), *Working papers in the theory of action* (pp. 111–161). New York: Free Press.
- Bales, R. F., & Strodtbeck, F. L. (1951). Phases in group problem solving. *Journal of Abnormal and Social Psychology*, 46, 485–495.
- Bandura, A. (1962). Social learning through imitation. In M. R. Jones (Ed.), *Nebraska symposium on motivation*. Lincoln: University of Nebraska Press.
- Baron, R. S., & Kerr, N. L. (2003). *Group process, group decision, group action* (2nd ed.). Buckingham, UK: Open University Press.
- Barrick, M. R., & Alexander, R. A. (1987). A review of quality circle efficacy and the existence of positive-findings bias. *Personnel Psychology*, 40, 579–592.
- Bartholomew, K., Henderson, A. J. Z., & Marcia, J. (2000). Coding semistructured interviews in social psychological research. In H. Reis & C. Judd (Eds.), *Research methods in social psychology: A handbook* (pp. 286–312). New York: Cambridge University Press.
- Betts, K. R., & Hinsz, V. B. (2010). Collaborative group memory: Processes, performance, and techniques for improvement. *Social and Personality Psychology Compass*, 4(2), 119–130.

- Borgatta, E. F., & Crowther, B. (1965). *A workbook for the study of social interaction processes*. Chicago: Rand-McNally.
- Bray, R., Kerr, N. L., & Atkin, R. (1978). Effects of group size, problem difficulty, and sex on group performance and member reactions. *Journal of Personality and Social Psychology*, 36, 1224–1240.
- Brewer, M. B., & Kramer, R. M. (1986). Choice behavior in social dilemmas: Effects of social identity, group size, and decision framing. *Journal of Personality and Social Psychology*, 50, 543–549.
- Bristol, T., & Fern, E. F. (1996). Exploring the atmosphere created by focus group interviews: Comparing consumers' feelings across qualitative techniques. *Journal of the Market Research Society*, 38, 185–195.
- Bristol, T., & Fern, E. F. (2003). The effects of interaction on consumers' attitudes in focus groups. *Psychology & Marketing*, 20(5), 433–454.
- Brown, B. B. (1968). *Delphi process: A methodology used for the elicitation of opinions of experts*. Santa Monica, CA: Rand Corporation.
- Buss, D. M. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences*, 12, 1–49.
- Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science*, 3, 14–25.
- Campbell, D. T. (1969). Definitional versus multiple operationism. *Et al.*, 2, 14–17.
- Cannell, C. F., & Kahn, R. (1968). Interviewing. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology*, Vol. 2 (2nd ed., pp. 526–595). Reading, MA: Addison Wesley.
- Chance, J. E., Goldstein, A. G., & McBride, L. (1975). Differential experience and recognition memory for faces. *Journal of Social Psychology*, 97, 243–253.
- Choi, J., & Bowles, S. (2007). The coevolution of parochial altruism and war. *Science*, 318(5850), 636–640.
- Chute, D. L. (1993). MacLaboratory for psychology: Successes, failures, economics, and outcomes over its decade of development. *Behavior Research Methods, Instruments, & Computers*, 25, 180–188.

- Cillessen, A. H. N. (2009). Sociometric methods. In K. H. Rubin, W. M. Bukowski, & B. Laursen (Eds.), *Handbook of peer interactions, relationships, and groups: Social, emotional, and personality development in context* (pp. 82–99). New York: Guilford Press.
- Coch, L., & French, J. R. (1948). Overcoming resistance to change. *Human Relations*, 1, 512–532.
- Cohen, J., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers*, 25, 257–271.
- Cohn, J. F., & Sayette, M. A. (2010). Spontaneous facial expression in a small group can be automatically measured: An initial demonstration. *Behavior Research Methods*, 42(4), 1079–1086.
- Curseu, P. L., & Wessel, I. (2008). How do virtual teams process information? A literature review and implications for management. *Journal of Managerial Psychology*, 23(6), 628–652.
- Dabbs, J. M., & Swiedler, T. C. (1983). Group AVTA: A microcomputer system for group voice chronography. *Behavior Research Methods and Instrumentation*, 15, 79–84.
- Dalkey, N. C. (1968). *Experiments in group prediction*. Santa Monica, CA: Rand Corporation.
- Dalkey, N. C. (1969–1970). *The Delphi method* (#s RM-5888-PR, RM-5957-PR, RM-6115-PR, RM-6118-PR). Santa Monica, CA: Rand Corporation.
- Dalkey, N. C., & Rourke, D. L. (1971). *Experimental assessment of Delphi procedures with group value judgments*. Santa Monica, CA: Rand Corporation.
- Davis, J. H., & Kerr, N. L. (1986). Thought experiments and the problem of sparse data in small-group performance research. In P. Goodman (Ed.), *Designing effective work groups*. New York: Jossey-Bass.
- Davis, J. H., Kerr, N. L., Atkin, R., Holt, R., & Meek, D. (1975). The decision processes of 6-and 12-person mock juries assigned unanimous and 2/3 majority rules. *Journal of Personality and Social Psychology*, 32, 1–14.

- Dawes, R. M., McTavish, J., & Shaklee, H. (1977). Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *Journal of Personality and Social Psychology*, 35, 1–11.
- Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). *Group techniques for program planning*. Glenview, IL: Scott, Foresman.
- Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1986). *Group techniques for program planning: A guide to nominal group and Delphi processes*. Middleton, WI: Greenbriar.
- Dennis, A. R., & Valacich, J. S. (1993). Computer brainstorming: More heads are better than one. *Journal of Applied Psychology*, 78, 531–537.
- Dennis, A. R., & Williams, M. L. (2003). Electronic brainstorming: Theory, research, and future directions. In P. Paulus & B. Nijstad (Eds.), *Group creativity: Innovation through collaboration* (pp. 160–178). New York: Oxford University Press.
- Deutsch, M., & Krauss, R. M. (1962). Studies of interpersonal bargaining. *Journal of Conflict Resolution*, 6, 52–76.
- Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology*, 53, 497–509.
- Diener, E., Lusk, R., DeFour, D., & Flax, R. (1980). Deindividuation: Effects of group size, density, number of observers, and group member similarity on self-consciousness and disinhibited behavior. *Journal of Personality and Social Psychology*, 39, 449–459.
- Dowling, K. L., & St. Louis, R. D. (2000). Asynchronous implementation of the nominal group technique: Is it effective? *Decision Support Systems*, 29(3), 229–248.
- Duggan, E. W. (2003). Generating systems requirements with facilitated group techniques. *Human-Computer Interaction*, 18(4), 373–394.
- Fern, E. F. (2001). *Advanced focus group research*. Thousand Oaks, CA: Sage.
- Festinger, L., Reicken, H. W., & Schachter, S. (1956). *When prophecy fails*. Minneapolis: University of Minnesota Press.
- Festinger, L., Schachter, S., & Back, K. (1950). *Social pressures in informal*

groups. New York: Harper.

- Filkins, J., Smith, C. M., & Tindale, R. S. (1998). The fairness of death qualified juries: A meta-analytic/simulation approach. In R. S. Tindale *et al.* (Eds.), *Social psychological applications to social issues: Applications of theory and research on small groups* (Vol. 4, pp. 153–176). New York: Plenum Press.
- Forsyth, D. R. (2010). *Group dynamics* (5th ed.). Belmont, CA: Wadsworth.
- Frankel, S. (1987). NGT +MDS: An adaptation of the nominal group technique for ill-structured problems. *Journal of Applied Behavioral Science*, 23, 543–551.
- Freedman, J. L., Klevansky, S., & Ehrlich, P. R. (1971). The effect of crowding on human task performance. *Journal of Applied Social Psychology*, 1, 7–25.
- Futoran, G. C., Kelly, J. R., & McGrath, J. E. (1989). TEMPO: A time-based system for analysis of group interaction process. *Basic and Applied Social Psychology*, 10(3), 211–232.
- Godwin, W. F., & Restle, F. (1974). The road to agreement: Subgroup pressures in small group consensus processes. *Journal of Personality and Social Psychology*, 30(4), 500–509.
- Guetzkow, H. (1968). Differentiation of roles in task oriented groups. In D. Cartwright & A. Zander (Eds.), *Group dynamics: Research and theory* (pp. 512–526). New York: Harper & Row.
- Gustafson, D. H., Shukla, R. K., Delbecq, A., & Walster, G. W. (1973). A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups, and nominal groups. *Organizational Behavior and Human Decision Processes*, 9, 280–291.
- Hackman, J. R., & Morris, C. G. (1975). Group tasks, group interaction process and group performance effectiveness: A review and proposed integration. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 8,). New York: Academic Press.
- Hannaford, P. L., Hans, V. P., & Munsterman, G. T. (2000). Permitting jury discussions during trial: Impact of the Arizona reform. *Law & Human Behavior*, 24(3), 359–382.
- Hänninen, L., & Pastell, M. (2009). CowLog: Open-source software for coding

- behaviors from digital video. *Behavior Research Methods*, 41(2), 472–476.
- Hansen, D., Shneiderman, B., & Smith, M. A. (2011). *Analyzing social media networks with NodeXL: Insights from a connected world*. Burlington, MA: Morgan Kauffman.
- Hare, A. P. (1976). *Handbook of small group research* (2nd ed.). New York: Free Press.
- Harvey, N., Harrieta, C., & Fischer, I. (2000). Using advice and assessing its quality. *Organizational Behavior and Human Decision Processes*, 81(2), 252–273.
- Hastie, R., Penrod, S., & Pennington, N. (1983). *Inside the jury*. Cambridge, MA: Harvard University Press.
- Haxby, J., Parasuraman, R., LaLonde, F., & Abboud, H. (1993). SuperLab: General-purpose Macintosh software for human experimental psychology and psychological testing. *Behavior Research Methods, Instruments, & Computers*, 25, 400–405.
- Haythorn, W. (1953). The influence of individual members on the characteristics of small groups. *Journal of Abnormal and Social Psychology*, 48, 276–284.
- Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.
- Helmer, O. (1966). *The use of the Delphi technique in problems of educational innovations*. Santa Monica, CA: Rand Corporation.
- Henrich, T. R., & Greene, T. J. (1991). Using the nominal group technique to elicit roadblocks to MRP II implementation. *Computers & Industrial Engineering*, 21, 335–338.
- Hertel, G., Geister, S., & Konradt, U. (2005). Managing virtual teams: A review of current empirical research. *Human Resource Management Review*, 15, 69–95.
- Hill, G. W. (1982). Group versus individual performance: Are $n+1$ heads better than one? *Psychological Bulletin*, 91, 517–539.
- Hoffman, L. R., & Maier, N. R. F. (1964). Valence in the adoption of solutions by problem-solving groups: Concept, method, and results. *The Journal of Abnormal and Social Psychology*, 69(3), 264–271.

- Hoffman, R., & MacDonald, J. (1993). Using HyperCard and Apple events in a network environment: Collecting data from simultaneous experimental sessions. *Behavior Research Methods, Instruments, & Computers*, 25, 114–126.
- Hollingshead, A. B., McGrath, J. E., & O'Connor, K. M. (1993). Group task performance and communication technology: A longitudinal study of computer-mediated versus face-to-face work groups. *Small Group Research*, 24, 307–333.
- Hunt, S. M. J. (1994). MacProbe: A Macintosh-based experimenter's workstation for the cognitive sciences. *Behavior Research Methods, Instruments, & Computers*, 26, 345–351.
- Hutchins, D. (1985). *Quality circles handbook*. London: Pitman.
- Hyman, H. H. (1978). *Interviewing in social research*. Chicago: University of Chicago Press.
- Ingle, S. (1982). *Quality circles master guide*. Englewood Cliffs, NJ: Prentice-Hall.
- Jarvis, B. (2011). Personal communication, October 20.
- Judd, C. M., & Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. *Journal of Personality and Social Psychology*, 54, 778–788.
- Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). *Research methods in social relations* (6th ed.). Fort Worth, TX: Harcourt, Brace, Jovanovich.
- Kafer, N. F. (1976). A sociometric method for identifying group boundaries. *Journal of Experimental Education*, 45, 71–74.
- Kameda, T., Takezawa, M., & Hastie, R. (2003). The logic of social sharing: An evolutionary game analysis of adaptive norm development. *Personality and Social Psychology Review*, 7(1), 2–19.
- Katz, N., Lazer, D., Arrow, H., & Contractor, N. (2005). The network perspective on small groups: Theory and research. In M. S. Poole & A. B. Hollingshead (Eds.), *Theories of small groups: Interdisciplinary perspectives* (pp. 277–312). Thousand Oaks, CA: Sage Publications
- Keeney, S., McKenna, H., & Hasson, F. (2011). *The Delphi technique in nursing*

and health research. Oxford: Blackwell-Wiley.

- Kelley, H. H., Condry, J. C., Dahlke, A., & Hill, A. (1965). Collective behavior in a simulated panic situation. *Journal of Experimental Social Psychology*, 1, 20–54.
- Kent, R. N., & McGrath, J. E. (1969). Task and group characteristics as factors influencing group performance. *Journal of Experimental Social Psychology*, 5(4), 429–440.
- Kerr, N. L. (1981). Social transition schemes: Charting the group's road to agreement. *Journal of Personality and Social Psychology*, 41, 684–702.
- Kerr, N. L., & Bray, R. M. (2005). Simulation, realism, and the study of the jury. In N. Brewer & K. Williams (Eds.), *Psychology and law: An empirical perspective* (pp. 322–360). New York: Guilford.
- Kerr, N. L., & Huang, J. Y. (1986). How much difference does one juror make in jury deliberation. *Personality and Social Psychology Bulletin*, 12, 325–343.
- Kerr, N. L., & MacCoun, R. (1985). Effects of jury size and polling method on the process and product of jury decision making. *Journal of Personality and Social Psychology*, 48, 349–363.
- Kerr, N. L., MacCoun, R., & Kramer, G. P. (1996) Bias in judgment: Comparing individuals and groups. *Psychological Review*, 103, 687–719.
- Kerr, N. L., & Tindale, R. S. (2011). Group-based forecasting? A social psychological analysis. Special Issue on “Enhancing group-based judgmental forecasting: Processes and priorities.” *International Journal of Forecasting*, 27(1), 14–40.
- Keyton, J., & Beck, S. J. (2009). The influential role of relational messages in group interaction. *Group Dynamics: Theory, Research, and Practice*, 13(1), 14–30.
- Kiesler, S. (1997). *Culture of the internet*. Mahwah, NJ: Erlbaum.
- Klein, K. J., & Kozlowski, S. W. J. (2000), *A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes*. San Francisco: Jossey-Bass.
- Knoke, D., & Kuklinski, J. H. (1982). *Network analysis*. Beverly Hills, CA: Sage.

- Knoke, D., & Yang, S. (2008). *Social network analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Komorita, S. S., & Chertkoff, J. M. (1973). A bargaining theory of coalition formation. *Psychological Review*, 80, 149–162.
- Korhonen, L. J. (1990). Nominal group technique. In M. W. Galbraith (Ed.), *Adult learning methods* (pp. 247–259). Melbourne, FL: Krieger.
- Kravitz, D. A., & Martin, B. (1986). Ringelmann rediscovered: The original article. *Journal of Personality and Social Psychology*, 50, 936–941.
- Krueger, R. A., & Casey, M. A. (2008). *Focus groups: A practical guide for applied research* (4th ed.). Thousand Oaks, CA: Sage.
- Lago, P. P., Beruvides, M. G., Jian, J.-Y., Canto, A. Y., Sandoval, A., & Taraban, R. (2007). Structuring group decision making in a web-based environment by using the nominal group technique, *Computers & Industrial Engineering*, 52(2), 277–295.
- Lampkin, E. C. (1972). Effects of n-dominance and group composition on task efficiency in laboratory triads. *Organizational Behavior & Human Performance*, 7(2), 189–202.
- Larson, J. R. (2007). Deep diversity and strong synergy: Modeling the impact of variability in members problem-solving strategies on group problem-solving performance. *Small Group Research*, 38(3), 413–436.
- Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* New York: Appleton-Century-Crofts.
- Latané, B., & L'Herrou, T. (1996). Spatial clustering in the conformity game: Dynamic social impact in electronic groups. *Journal of Personality and Social Psychology*, 70, 1218–1230.
- Latané, B., Williams, K. D., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, 37, 822–832.
- Laughlin, P. R. (1996). Group decision making and collective induction. In E. Witte & J. H. Davis (Eds.), *Understanding group behavior* (Vol. 1, pp. 61–80). Mahwah, NJ: Erlbaum.
- Laughlin, P. R., Branch, L. G., & Johnson, H. H. (1969). Individual versus

- triadic performance on a unidimensional complementary task as a function of initial ability level. *Journal of Personality and Social Psychology*, 12(2), 144–150.
- Laughlin, P. R., Hatch, E. C., Silver, J. S., & Boh, L. (2006). Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *Journal of Personality and Social Psychology*, 90, 644–651.
- Laughlin, P. R., & Johnson, H. H. (1966). Groups and individual performance on a complementary task as a function of initial ability level. *Journal of Experimental Social Psychology*, 2, 407–414.
- Leavitt, H. J. (1951). Some effects of certain communication patterns on group performance. *Journal of Abnormal and Social Psychology*, 46, 38–50.
- Leggett-Dugosh, K., Paulus, P. B., Roland, E. J., & Yang, H. C. (2000). Cognitive stimulation in brainstorming. *Journal of Personality and Social Psychology*, 79(5), 722–735.
- Levin, M. L. (1976). Displaying sociometric structures: An application of interactive computer graphics for instruction and analysis. *Simulation and Games*, 7, 295–310.
- Levine, J. M. (1978). GROUPCOM: A computer program for investigating social processes in small groups. *Behavior Research Methods, Instruments, & Computers*, 10, 191–195.
- Levine, J. M., & Moreland, R. L. (1990). Progress in small group research. *Annual Review of Psychology*, 41, 585–634.
- Levine, J. M., & Moreland, R. L. (1997). Small groups. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 416–459). Boston: McGraw-Hill.
- Lewin, K. (1953). Studies in group decision. In D. Cartwright & A. Zander (Eds.), *Group dynamics: Research and theory*. Evanston, IL: Row, Peterson.
- Lewin, K., Lippett, R., & White, R. (1939). Patterns of aggressive behavior in experimentally created “social climates.” *Journal of Social Psychology*, 10, 271–299.
- Li, J., Seu, J., Evens, M., Michael, J., & Rovick, A. (1992). Computer dialogue system (CDS): A system for capturing computer-mediated dialogue. *Behavior*

Research Methods, Instruments, & Computers, 24, 535–540.

Liamputpong, P. (2011). *Focus group methodology: Principle and practice*. Thousand Oaks, CA: Sage.

Lindzey, G., & Borgatta, E. F. (1954). Sociometric measurement. In G. Lindzey (Ed.), *Handbook of social psychology* (pp. 405–448). Cambridge, MA: Addison-Wesley.

MacLin, O. H., & MacLin, M. K. (2005). Coding observational data: A software solution. *Behavior Research Methods*, 37(2), 224–231.

Martino, J. P. (1983). *Technological forecasting for decision making*. New York: Elsevier Science Publishing.

Matsatsinis, N. F., Grigoroudis, E., & Samaras, A. (2005). Aggregation and disaggregation of preferences for collective decision-making. *Group Decision and Negotiation*, 14(3), 217–232.

Mayo, E. (1933). *The human problems of an industrial civilization*. Cambridge, MA: Harvard University Press.

McGrath, J. E. (1984). *Groups: Interaction and performance*. Englewood Cliffs, NJ: Prentice-Hall.

McGrath, J. E., & Altermatt, T. W. (2001). Observation and analysis of group interaction over time: Some methodological and strategic choices. In M. Hogg & S. Tindale (Eds.), *Blackwell handbook in social psychology* (Vol. 3, pp. 525–556). Cambridge, MA: Blackwell.

McGrath, J. E., & Altman, I. (1966). *Small group research: A synthesis and critique of the field*. New York: Holt.

McGrath, J. E., & Hollingshead, A. B. (1994). *Groups interacting with technology*. Thousand Oaks, CA: Sage.

McGrath, J. E., & Tschan, F. (2004). *Temporal matters in social psychology: Examining the role of time in the lives of groups and individuals*. Washington, DC: American Psychological Association Publications.

McGuire, T., Kiesler, S., & Siegel, S. (1987). Group and computer-mediated discussion effects in risk decision making. *Journal of Personality and Social Psychology*, 52, 917–930.

- Messick, D. M., Wilke, H. A. M., Brewer, M. B., Kramer, R. M., Zemke, P. E., & Lui, J. (1983). Individual adaptations and structural change as solutions to social dilemmas. *Journal of Personality and Social Psychology*, 44, 294–309.
- Milgram, S. (1974). *Obedience to authority: An experimental view*. New York: Harper & Row.
- Moreland, R. L. (1985). Social categorization and the assimilation of “new” group members. *Journal of Personality and Social Psychology*, 48, 1173–1190.
- Moreland, R. L., Fetterman, J. D., Flagg, J. J., & Swanenburg, K. L. (2010). Behavioral assessment practices among social psychologists who study small groups. In C. R. Agnew, D. E. Carlston, W. G. Graziano, & J. R. Kelly (Eds.), *Then a miracle occurs: Focusing on behavior in social psychological theory and research* (pp. 28–56). New York: Oxford University Press.
- Moreland, R. L., Hogg, M. A., & Hains, S. C. (1994). Back to the future: Social psychological research on groups. *Journal of Experimental Social Psychology*, 30, 527–555.
- Moreland, R. L., & Levine, J. M. (1982). Socialization in small groups: Temporal changes in individual-group relations. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 15, pp. 137–192). New York: Academic Press.
- Moreno, J. L. (1953). *Who shall survive?* (Rev. ed.). Beacon, NY: Beacon House.
- Mullen, B., Johnson, C., & Salas, E. (1991). Productivity loss in brainstorming groups: A meta-analytic integration. *Basic and Applied Social Psychology*, 12, 3–24.
- Newcomb, T. M. (1961). *The acquaintance process*. New York: Holt.
- Nijstad, B. A. (2009). *Group performance*. New York: Psychology Press.
- Nijstad, B. A., Stroebe, W., & Lodewijckx, H. F. M. (2002). Cognitive stimulation and interference in groups: Exposure effects in an idea generation task. *Journal of Experimental Social Psychology*, 38(6), 535–544.
- Nijstad, B. A., Stroebe, W., & Lodewijckx, H. F. M. (2006). The illusion of group productivity: A reduction of failures explanation. *European Journal of Social*

Psychology, 36(1), 31–48.

- Noldus, L. P. J. J., Trienes, R. J. H., Hendriksen, A. H. M., Jansen, H., & Jansen, R. G. (2000). The observer video-pro: New software for the collection, management, and presentation of time-structured data from videotapes and digital media files. *Behavior Research Methods, Instruments & Computers*, 32(1), 197–206.
- Noma, E., & Smith, D. R. (1978). SHED: A FORTRAN IV program for the analysis of small group sociometric structure. *Behavior Research Methods & Instrumentation*, 10, 60–62.
- Northway, M. L. (1967). *A primer of sociometry* (2nd ed.). Toronto, ON: University of Toronto Press.
- Offner, A. K., Kramer, T. J., & Winter, J. P. (1996). The effects of facilitation, recording, and pauses on group brainstorming. *Small Group Research*, 27(2), 283–98.
- Olson, J. S., Olson, G., Storreston, M., & Carter, M. (1994). Groupwork close up: A comparison of the group design process with and without a simple group editor. *ACM Transactions on Information Systems*, 11, 321–348.
- Osborn, A. F. (1957). *Applied imagination* (Rev. ed.). New York: Scribner.
- Oxley, N. L., Dzindolet, M. T., & Paulus, P. B. (1996). The effects of facilitators on the performance of brainstorming groups. *Journal of Social Behavior & Personality*, 11(4), 633–646.
- Park, J. (1991). Estimating success rates of quality circle programs: Public and private experiences. *Public Administration Quarterly*, 15(1), 133–146.
- Park, S., & Golembiewski, R. T. (1991). An examination of the determinants of successful QC programs: Testing the influence of eleven situational features. *Organization Development Journal*, 9(4), 38–49.
- Parson, T., & Bales, R. F. (1955). *Family, socialization and interaction process*. Glencoe, IL: Free Press.
- Paul, S., Haseman, W. D., & Ramamurthy, K. (2004). Collective memory support and cognitive-conflict group decision-making: An experimental investigation. *Decision Support Systems*, 36(3), 261–281.
- Paulus, P. B., & Dzindolet, M. T. (1993). Social influence processes in group

- brainstorming. *Journal of Personality and Social Psychology*, 64, 575–586.
- Pereira, G. M., & Osburn, H. G. (2007). Effects of participation in decision making on performance and employee attitudes: A quality circles meta-analysis. *Journal of Business and Psychology*, 22(2), 145–153.
- Philipsen, G., Mulac, A., & Dietrich, D. (1979). The effects of social interaction on group idea generation. *Communication Monographs*, 46, 119–125.
- Rapoport, A. (1967). Optimal policies for the prisoner's dilemma game. *Psychological Review*, 74, 136–148.
- Reis, H. T. (1983). The promise of naturalistic methods. *New Directions for Methodology of Social & Behavioral Science*, 15, 1–4.
- Reis, H. T., & Stiller, J. (1992). Publication trends in JPSP: A three-decade review. *Personality and Social Psychology Bulletin*, 18, 465–472.
- Ren, Y., Carley, K. M., & Argote, L. (2006). The contingent effects of transactive memory: When is it more beneficial to know what others know? *Management Science*, 52(5), 671–682.
- Roethlisberger, F. J., & Dickson, W. J. (1939). *Management and the worker*. Cambridge, MA: Harvard University Press.
- Rohrbaugh, J. (1979). Improving the quality of group judgment: Social judgment analysis and the Delphi technique. *Organizational Behavior and Human Performance*, 24, 73–92.
- Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting*, 15, 353–375.
- Rowe, G., & Wright, G. (2001): Expert opinions in forecasting: Role of the Delphi technique. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook of researchers and practitioners* (pp. 125–144). Boston: Kluwer Academic Publishers.
- Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, 5(1), 2–14.
- Ruback, R. B., Dabbs, J. M., & Hopper, C. H. (1984). The process of brainstorming: An analysis with individual and group vocal parameters. *Journal of Personality and Social Psychology*, 47, 558–567.

- Runkle, P., & McGrath, J. E. (1972). *Research on human behavior: A systematic guide to method*. New York: Holt.
- Sackman, H. (1975). *Delphi critique: Expert opinion, forecasting, and group process*. Lexington, MA: Lexington Books.
- Sanderson, P. M. (1994). Handling complex real-world data with two cognitive engineering tools: COGENT and MacSHAPA. *Behavior Research Methods, Instruments, & Computers*, 26, 117–124.
- Sandys, M., & Dillehay, R. C. (1995). First-ballot votes, predeliberation dispositions, and final verdicts in jury trials. *Law and Human Behavior*, 19, 175–195.
- Schachter, S. (1951). Deviation, rejection, and communication. *Journal of Abnormal and Social Psychology*, 46, 190–207.
- Schachter, S. (1959). *The psychology of affiliation*. Stanford, CA: Stanford University Press.
- Schachter, S., Ellertson, N., McBride, D., & Gregory, D. (1951). An experimental study of cohesiveness and productivity. *Human Relations*, 4, 229–238.
- Schneider, W. (1991). Equipment is cheap, but the field must develop and support common software for psychological research. *Behavior Research Methods, Instruments, & Computers*, 23, 114–116.
- Scott, J. (2000). *Social network analysis: A handbook*. London: Sage.
- Scott, J., & Carrington, P. J. (2011). *The Sage handbook of social network analysis*. Thousand Oaks, CA: Sage.
- Seal, D. W., Bogart, L. M., & Ehrhardt, A. A. (1998). Small group dynamics: The utility of focus group discussions as a research method. *Group Dynamics: Theory, Research, and Practice*, 2(4), 253–266.
- Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51(3), 515–530.
- Sherif, M. (1936). *The psychology of social norms*. New York: Harper.
- Sherif, M. (1966). *In common predicament: Social psychology of intergroup*

conflict and cooperation. New York: Houghton Mifflin.

Sherif, M., Harvey, O. J., White, B., Hood, W., & Sherif, C. (1961). *Intergroup conflict and cooperation*. Norman, OK: Institute of Group Relations.

Sherwin, R. G. (1975). Structural balance and the sociomatrix: Finding triadic valence structures in signed adjacency matrices. *Human Relations*, 28, 175–189.

Siegal, S., & Fouraker, L. E. (1960). *Bargaining and group decision making: Experiments in bilateral monopoly*. New York: McGraw-Hill.

Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, 11(1), 1–18.

Snizek, J. A. (1992). Groups under uncertainty: An examination of confidence in group decision making. *Organizational Behavior and Human Decision Processes*, 52(1), 124–155.

Snizek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes*, 62(2), 159–174.

Sommers, R. (1959). Studies in personal space. *Sociometry*, 22, 247–260.

Sproull, L., & Kiesler, S. B. (1991). *Connections: New ways of working in the networked organization*. Cambridge, MA: MIT Press.

Stasser, G., Kerr, N. L., & Davis, J. H. (1989). Influence processes and consensus models in decision-making groups. In P. Paulus (Ed.), *Psychology of group influence* (2nd ed., pp. 279–326). Hillsdale, NJ: Lawrence Erlbaum.

Stasser, G., & Taylor, L. (1991). Speaking turns in face-to-face discussion. *Journal of Personality and Social Psychology*, 60, 675–684.

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during group discussion. *Journal of Personality and Social Psychology*, 48, 1467–1478.

Steel, R. P., & Shane, G. S. (1986). Evaluation research on quality circles: Technical and analytical implications. *Human Relations*, 39, 449–468.

Steiner, I. (1972). *Group process and productivity*. New York: Academic Press.

- Steiner, I. (1974). Whatever happened to the group in social psychology? *Journal of Experimental Social Psychology*, 10, 94–108.
- Steiner, I. (1986). Paradigms and groups. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 251–292). Orlando, FL: Academic Press.
- Stephan, F. F., & Mishler, E. G. (1952). The distribution of participation in small groups: An exponential approximation. *American Sociological Review*, 17, 598–608.
- Stephenson, B. Y., Michaelsen, L. K., & Franklin, S. G. (1982). An empirical test of the nominal group technique in state solar energy planning. *Group & Organization Studies*, 7, 320–334.
- Stewart, D. W., Shamdasani, P. N., & Rook, D. (2006). *Focus groups: Theory and practice* (2nd ed.). Thousand Oaks, CA: Sage.
- Stroebe, W., & Diehl, M. (1994). Why groups are less effective than their members: On productivity losses in idea generating groups. *European Review of Social Psychology*, 5, 271–303.
- Stroebe, W., Diehl, M., & Abakoumkin, G. (1992). The illusion of group effectivity. *Personality and Social Psychology Bulletin*, 18, 643–650.
- Stumpf, S. A., Freedman, R. A., & Zand, D. E. (1979). Judgmental decisions: A study of interactions among group membership, group functioning, and the decision situation. *Academy of Management Journal*, 22, 765–782.
- Tajfel, H., Billig, M., Bundy, R., & Flament, C. (1971). Social categorization and intergroup behavior. *European Journal of Social Psychology*, 1, 149–178.
- Takagi, E. (1999). Solving social dilemmas is easy in a communal society: A computer simulation analysis. In M. Foddy, M. Smithson, S. Schneider, & M. Hogg (Eds.), *Resolving social dilemmas: Dynamic, structural, and intergroup aspects* (pp. 33–54). New York: Psychology Press.
- Tapp, J., & Walden, T. (1993). PROCODER: A professional tape control, coding, and analysis system for behavioral research using videotape. *Behavior Research Methods, Instruments, & Computers*, 25, 53–56.
- Tetlock, P. E. (1979). Identifying victims of groupthink from public statements of decision makers. *Journal of Personality and Social Psychology*, 37(8),

1314–1324.

- Tindale, R. S., & Nagao, D. H. (1986). An assessment of the potential utility of “Scientific Jury Selection”: A “thought experiment” approach. *Organizational Behavior and Human Decision Processes*, 37, 409–425.
- Tindale, R. S., & Vollrath, D. A. (1992). “Thought experiments” in applied social research. In F. B. Bryant, J. E. Edwards, R. S. Tindale, E. J. Posavac, L. Heath, E. Henderson, & Y. Suarez-Balcazar (Eds.), *Social psychological applications to social issues: Methodological issues in applied social research* (Vol. 2, pp. 219–238). New York: Plenum Press.
- Valacich, J. S., Dennis, A. R., & Connolly, T. (1994). Idea generation in computer-based groups: A new ending to an old story. *Organizational Behavior & Human Decision Processes*, 57, 448–467.
- Van de Ven, A. H. (1974). *Group decision making and effectiveness: An experimental study*. Kent, OH: Kent State University Press.
- Vaughan, J., & Yee, P. L. (1994). Using PsyScope for demonstrations and student-designed experiments in cognitive psychology courses. *Behavior Research Methods, Instruments, & Computers*, 26, 142–147.
- Vaughn, S., Schumm, J. S., & Sinagub, J. (1996). *Focus group interviews in education and psychology*. Thousand Oaks, CA: Sage.
- Wallack, M. A., Kogan, W., & Bem, D. J. (1962). Group influence in individual risk taking. *Journal of Abnormal and Social Psychology*, 65, 75–86.
- Walster, E., Walster, G., & Bersheid, E. (1978). *Equity: Theory and research*. Boston: Allyn & Bacon.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.
- Watanabe, Y., & Yamagishi, T. (1999). Emergence of strategies in a selective play environment with geographic mobility: A computer simulation. In M. Foddy, M. Smithson, S. Schneider, & M. Hogg (Eds.), *Resolving social dilemmas: Dynamic, structural, and intergroup aspects* (pp. 55–66). New York: Psychology Press.
- Weber, A. L., & Harvey, J. H. (Eds.). (1994). *Perspectives on close relationships*. Boston: Allyn & Bacon.

- Weick, K. E. (1985). Systematic observational methods. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (3rd ed., pp. 567–634). New York: Random House.
- Weingart, L. (1997). How did they do that? The ways and means of studying group processes. *Research in Organizational Behavior*, 19, 189–239.
- Weldon, E., & Weingart, L. R. (1993). Group goals and group performance. *British Journal of Social Psychology*, 32(4), 307–334.
- Wheelan, S. A. (1994). *Group processes: A developmental perspective*. Boston: Allyn & Bacon.
- Whyte, W. F. (1943). *Street corner society*. Chicago: University of Chicago Press.
- Wilson, J. P., Aronoff, J., & Messé, L. A. (1975). Social structure, member motivation, and productivity. *Journal of Personality and Social Psychology*, 32, 1094–1098.
- Wittenbaum, G. M., & Moreland, R. L. (2008). Small group research in social psychology: Topics and trends over time. *Social and Personality Psychology Compass*, 2, 187–208.
- Wolfe, C. (1992). Using Authorware Professional for developing courseware. *Behavior Research Methods, Instruments, & Computers*, 24, 273–276.
- Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83(2), 260–281.
- Zajonc, R. B. (1965). Social facilitation. *Science*, 149, 269–274.
- Zimmerman, P. H., Bolhuis, J. E., Willemsen, A., Meyer, E. S., & Noldus, L. P. J. J. (2009). The observer XT: A tool for the integration and synchronization of multimodal signals. *Behavior Research Methods*, 41(3), 731–735.

¹ Historically, the defining features of the small group have also been the focus of some debate (cf. Forsyth, 2009). As a way of demarking a set of research question and substantive phenomenon, we like McGrath's (1984) flexible, fuzzy-set definition of the group: “an aggregation of two or more people who are to

some degree in dynamic interrelation with one another” (p. 8). However, in the present context, we believe that the definition that we present here is both serviceable and consistent with most perspectives on group phenomena.

² We should note that we do not address all of the methodological issues that arise in the study of groups in this chapter. In particular, see Kenny & Kashy (Chapter 19 in this volume) for a detailed discussion of how to handle some of the statistical problems that arise in the analysis of group data, and see Klein and Kozlowski (2000) for an introduction to the conceptual and methodological problems that arise when studying collective phenomena at varying levels of analysis.

³ Note that consideration of observation-recording techniques is also relevant to some laboratory-based research, particularly the type, discussed more fully later, that has participants interact face to face, with few constraints on behavior.

Chapter ten Inducing and Measuring Emotion and Affect

Tips, Tricks, and Secrets

Karen S. Quigley, Kristen A. Lindquist, and and Lisa Feldman Barrett

Every person (barring those with a brain disorder) knows what it feels like to be moved by something – to feel energized or defeated, anxious or tranquil. Even without labeling these feelings, or being aware of them in an explicit way, such feelings exist as states of mind or can be observed in certain actions. In the Western views of the human mind that ground scientific psychology, such states are referred to as “emotional” or “affective” (as distinguished from “cognitive” or “perceptual”).¹ These two words – “emotion” and “affect” – have caused great confusion in the scientific literature because they are used by some authors to denote two different classes of phenomena, whereas others use these words interchangeably. In English, the word “affect” literally means “to produce a change,” whereas the word “emotion” derives from the French word “to stir up” and the Latin word “to move.” In psychological discourse, “affect” has sometimes been used to refer to free-floating feelings whereas “emotion” has referred to feelings in response to a specific triggering event (e.g., James, 1890). The word “affect” also has been used to refer to feelings that accompany emotions such as anger, sadness, fear, happiness, and so on, which are defined as physical states (e.g., Panksepp, 1998). “Affect” has been used as a general term to mean anything emotional (e.g., Davidson, Scherer, & Goldsmith, 2003), allowing researchers to talk about emotion in a theory-neutral way. And sometimes “affect” is used to refer to hedonic valence and arousal (e.g., Barrett & Russell, 1998; Russell, 2003) or to approach or avoidance action tendencies (e.g., Lang, Bradley, & Cuthbert, 1990) that are common to experiences and perceptions of emotion, as well as to refer to the motivating, engaging core of all mental states covering a range of psychological phenomena, including but not limited to emotion (Barrett & Bliss-Moreau, 2009b); in such cases, emotions are designated as discrete states of anger, fear, sadness, disgust, and happiness (plus a few others, depending on the theorist) in which affect is meaningfully linked to a situation in some causal way.

In this chapter, we review the typical methods that are used to create and measure the physical states and subjective feelings that researchers refer to as “affect” or “emotion,” keeping in mind the scientific distinction between these two constructions. We refer to “affect” as the properties of any mental state that can be described as pleasant or unpleasant with some degree of arousal (Barrett & Bliss-Moreau, 2009a; Russell & Barrett, 1999). These properties correspond to brain representations of some change in the core autonomic and hormonal systems of the body (whether or not such changes actually take place). There is no widely accepted operational definition of emotion. Sometimes writers describe emotion as coordinated packets of experiences, physiological changes, and behavior, but this is nonspecific because every waking moment of life there are coordinated changes of this sort. Furthermore, there remains tremendous debate over which mental states count as emotion versus which do not (e.g., Ortony & Turner, 1990). In this chapter we take a simple approach: an “emotion” is a mental state to which people assign a commonsense name (like anger, sadness, fear, disgust, happiness, and a handful of others like shame, guilt, pride, and so on); when someone uses an “emotion” label, it implies they have invoked conceptual knowledge about emotion to make sense of or to communicate their internal state. From our perspective, inducing emotion necessarily involves a change in affect (whereas changes in affect are not always transformed into emotions). This means that to make claims about emotion, it is necessary to ensure that findings do not simply reflect changes in valence or arousal. Furthermore, there are times when a scientist's intention to evoke an affective change in a participant produces an unexpected change in an emotion (e.g., showing a participant an image of a dying person, which evokes a memory of a family member who died recently).

With these considerations in mind, we very generally review the variety of induction methods and measurement techniques that are used most frequently in social and personality psychology. For more detailed treatments, see the *Handbook of Emotion Elicitation and Assessment* (Coan & Allen, 2007) and the *Handbook of Affective Sciences* (Davidson et al., 2003). We highlight novel points related to inducing affect or emotions as *experiences* or *states* and discuss the most serious challenges that researchers face, the most serious being that at times the intent is to measure changes in *emotion* when the measurement tools only permit inferences about *affect*. Currently, there is no strong empirical justification for using any single objective measurement, or profile of measurements (in the face, body, or brain) to indicate when a person is in a state of anger, or fear, or sadness, and so on. People do not always scowl in anger,

heart rate does not always go down in sadness, and people do not always freeze or run in fear. Reviews of the empirical literature have reached this conclusion again and again over the past hundred years (Lindquist, Siegel, Quigley, & Barrett, 2013). Yet it is possible to have a powerful and robust science of emotion, when induction methods are used judiciously and measurements are interpreted appropriately. This chapter is designed to help interested readers move forward in that direction.

Methods for Inducing Affective Changes, Including Emotions

We outline thirteen laboratory induction techniques that are the most frequently and successfully used laboratory-based inductions. A brief summary of each method is also presented in Table 10.1, including a description, prototypical references, and advantages and disadvantages of each method. For a more extensive Supplemental Table 10.1, see <http://www.affective-science.org/publications.shtml>. Because emotions are a subset of affective changes more generally, in principle, any stimulus that is used to induce affective changes (varying in hedonic valence and arousal) can also be used to evoke emotions (anger, sadness, fear, etc.) and vice versa, depending on the instructions given to the participant at encoding. Although we summarize methods typically used in the scientific literature for evoking affect more generally, and emotion more specifically, from our point of view it is possible to evoke an emotion whenever a stimulus or the context elicits conceptual knowledge about emotion (or when a perceiver is prompted to categorize a response as emotional either explicitly or implicitly using emotion words; we say more about this latter issue in the section on measuring emotion). Thus an emotion can be evoked, even when the experimenter's intent is to evoke affect. Conversely, when such conceptualization is prevented, then a stimulus is likely to evoke an affective response (even when the experimenter's intent is to evoke emotion).

Table 10.1. *Affect and Emotion Induction Techniques Including Methods, Exemplar References, Advantages, and Disadvantages*

Laboratory Inductions	Representative Stimulus Sets and References	Advantages	Disadvantages	Effect Size (g)
Films*	Methods and typical stimulus sets: (Gross & Levenson, 1995; Philippot, 1993; Schaefer, Nils, Sanchez, & Philippot, 2010).	Ease of presentation	Participant familiarity can introduce variability	.53–.66
Images*	Methods and typical stimulus sets: (Bradley et al., 2001; Lang et al., 1993) e.g., International Affective Picture System or IAPS; (Lang et al., 2008).	Ease of presentation	Slides do not sample all aspects of affective space	.58–1.03
Faces	Methods and typical stimulus sets: e.g., the Ekman and Friesen set (Ekman & Friesen, 1978), the Japanese and Caucasian Facial Expression of Emotion set (JACFEE; Matsumoto & Ekman, 1988); the Montreal Set of Facial Displays for Emotion (Beaupré & Hess, 2005).	Ease of presentation	Most faces are used in studies of emotion perception. It is not clear whether faces shift feelings or prime concepts.	n/a
Sounds/ Voices	Methods and typical stimulus sets: e.g., International Affective Digitized Sounds (IADS) (Bradley & Lang, 2007); Sounds can be affective because of their representational content (e.g., buzzing bees), because their acoustical properties make them intrinsically affective (e.g., sirens), human voices can speak neutral words or sentences with an affective tone, or prosody as in (Banse & Scherer, 1996; Bliss-Moreau, Owren, & Barrett, 2010), or stimuli can be nonlinguistic emotional utterances (e.g., grunting in anger; Simon-Thomas et al., 2009); or naturalistic (e.g., pilots speaking during dangerous flights). The latter have limitations (see Scherer, 2003), including that they confound emotional semantic content with prosody.	Ease of presentation	Most prosody stimuli are used in studies of emotion perception; sounds with acoustical properties that act directly on the nervous system (e.g., sirens) shift feelings; sounds with representational content (e.g., the sound of bees, affective prosody) might prime concepts.	n/a
Music*	Methods and typical stimulus sets: The Continuous Music Technique (Eich & Metcalfe, 1989) pairs classical music (with no explicit semantic content) with imagined events (either hypothetical or autobiographical) with the intent of intensifying feelings.	Music can be played in the background to keep evocative states elevated throughout an experiment	Music does not reliably induce specific discrete emotions (e.g., anger vs. anxiety) although it can induce valence effects (positive vs. negative vs. neutral)	.41–.65

Laboratory Inductions	Representative Stimulus Sets and References	Advantages	Disadvantages	Effect Size (g)
Imagery and Recall*	<p>Methods and typical stimulus sets: Open-ended imagery instructions are used in the Continuous Music Technique (see above) or scripts can be used. In the Scenario Immersion Technique, participants read (or hear) embodied scenarios and experience a narrative as it unfolds (e.g., Wilson-Mendenhall, Barrett, Simmons, & Barsalou, 2011). Another imagery approach involves gathering autobiographical details from the participant and then constructing idiographic narratives (Olatunji, Babson, Smith, Feldner, & Connolly, 2009). This differs from true autobiographical recall because the scenarios are constructed by the researchers into a structured narrative. The Velten technique (Velten, 1968) is a form of guided imagery; participants are given statements describing positive or negative self-evaluations and asked to imagine situations that apply to them (e.g., Carter et al., 2002). Recall and Velten had equal efficacy to other imagery approaches in the Lench et al. (2011) meta-analysis (table 1).</p>	Ecologically valid; content can be idiographically manipulated	Participants vary in the ability to engage in mental imagery which will increase variability	.42–.61 (Imagination) .39–.51 (Autobiographical recall)
Words	<p>Methods and typical stimulus sets: Typically, valenced words are used in evaluative priming paradigms where subliminally presenting a negative word (e.g., “murder”) prior to a same-valenced object (e.g., a snake) speeds a participant’s latency to respond (Ferguson, Bargh, & Nayak, 2005); exemplar words can be found in the Affective Norms for English Words set (ANEW); like IAPS images, ANEW words have been rated for valence and arousal (Bradley & Lang, 1999) and discrete emotions (Stevenson, Mikels, & James, 2007).</p>	Ease of presentation	Most words are used in studies of evaluative priming; as induction stimuli, it is not clear whether words shift feelings or prime concepts	.02–.49

(continued)

Laboratory Inductions	Representative Stimulus Sets and References	Advantages	Disadvantages	Effect Size (g)
Bodily Movements and Postures	Methods and typical stimulus sets: e.g., facial muscle manipulation using a pen held in the teeth vs. lips (Strack et al., 1988). Nonfacial bodily movements include asking participants to use approach or avoidance-related flexion or extension-based muscle movements or head movements (see example outcomes in the supplemental version of this table), or take postures suggestive of a particular emotion state (e.g., Duclos et al., 1989, Study 2, postures of fear, sadness, or anger) or a gross change in posture (e.g., slumping; Stepper & Strack, 1993, Study 1).	May be a relatively implicit manner of shifting feelings	Researchers must present a good cover story to prevent demand characteristics	.34–.60
Peripheral physiological manipulations	Methods: e.g., injections of epinephrine (Schacter & Singer, 1962); exercise: (Ekkekakis et al., 2011); oxytocin: (for review, see Norman, Hawkey et al., 2011); botox into facial muscles: (Davis et al., 2010).	Acts directly on peripheral physiological systems and can be quite potent	Requires expertise to administer safely	n/a
Confederates	Methods: e.g., using a confederate to induce anger (Cohen et al., 1996); using a confederate to induce jealousy in participants by forming a bond with one participant, and then choosing to work with another participant in a subsequent task (DeSteno, Valdesolo, & Bartlett, 2006).	Ecologically valid	A good cover story is critical so participants cannot guess the confederate's role; requires extensive planning at design and implementation	.37–.54
Motivated performance	Methods and typical stimulus sets: e.g., the Trier Social Stress Task (TSST; Kirschbaum et al., 1993); a variation of the TSST is used to induce either unpleasant feelings of social rejection or pleasant social approval by varying whether audience members were unsupportive or supportive (Akinola & Mendes, 2008) or altered their speech prosody to induce shame (negative feedback in a warm tone) or anger (negative feedback in a condescending tone; Kassam & Mendes, 2013).	Ecologically valid	Requires screening to ensure that participants who will find the task too evocative do not participate (e.g., participants with social anxiety)	n/a

Laboratory Inductions	Representative Stimulus Sets and References	Advantages	Disadvantages	Effect Size (g)
Virtual reality	Methods and typical stimulus sets: e.g., experience of a virtual park (Riva et al., 2007); virtual reality exposure therapy for treating anxiety (e.g., Michaliszyn, Marchand, Bouchard, Martel, & Poirier-Bisson, 2010). <u>Virtual park:</u> Participants explored a virtual urban park (with trees, benches, lights, walking paths). Affect was induced by changing background music, lighting, shadows, and the presence of other people. <u>Virtual reality exposure therapy:</u> Virtual reality used to induce affect in the treatment of anxiety, particularly for exposure therapy. A meta-analysis showed effects comparable to clinical <i>in vivo</i> exposures (Powers & Emmelkamp, 2008).	Advantages include good experimental control and repeatability, potential for enhanced believability of manipulations, and ability to present well designed social manipulations to heterogeneous samples to enhance generalizability (Blascovich & Bailenson, 2011)	The downside of this technology is the cost and the need for some degree of technological sophistication, especially in programming (Blascovich & Bailenson, 2011)	n/a
Physically Real Stimuli (including Experience Sampling)	Methods and typical stimulus sets: Real stimuli include: spiders/snakes to test avoidance of feared objects (Teachman, 2007), sky diving and mountaineering in extreme sports enthusiasts (Castanier, LeScanff, & Woodman, 2011), foods or other substances to induce disgust or pleasure (Jabbi, Swart, & Keyers, 2007), nociceptive stimuli (Lovallo et al., 1985), and chemosensory (i.e., odor) stimuli (for review, see Yesharun & Sobel, 2010).	Ecologically valid and impactful	More difficult to administer and more idiographic variation, and thus reduced experimental control. They are also often more costly and time-consuming to use.	n/a

Note: Effect sizes shown (where applicable) are the 95% confidence interval provided in Lench *et al.* (2011). Effect sizes are Hedges' g (.2 considered a small effect, .5 considered medium, and .8 considered large).

* Inductions with effect sizes greater than .5 in Table 1 of Lench *et al.* (2011).

Films

The entertainment industry knows that people will pay a lot of money to see a movie, precisely because movies powerfully influence momentary experience. Several scholarly works have proposed theoretical frameworks for understanding

how films evoke affective and emotional changes (e.g., Allen & Smith, 1997; Tan, 2000; see Table 10.1 for references to film clip sets). Films are easy to use. In the typical film-based induction, participants are seated in front of a blank television or computer screen and asked to relax for a 1–3-minute baseline period after which they view a film clip for 2–5 minutes on average. A downside is that participants will vary in their familiarity with the movie clips, which introduces variability as error variance (because familiarity can influence potency). Manipulation checks (after the film) should be performed with caution because presenting adjectives to a participant and having the participant rate his or her state with those words have the potential to transform an affective state into an emotional one, or to change one emotional state to another, over and above the impact of the induction itself. If attempting to induce a change in affect, consider using an affect-based rating scale like the Self-Assessment Manikin (SAM; Bradley & Lang, 1994) or a two-dimensional affect grid (Russell, Weiss, & Mendelsohn, 1989) as a manipulation check. With both measures, it is important to clearly define arousal (high vs. low activation), as this property is not identical to the intensity of experience, although the two are often confused (Kuppens, Tuerlinckx, Russell & Barrett, 2012). Also, keep in mind that any rating has the potential to reduce the intensity of the induced change (e.g., Lieberman, Eisenberger, Crockett, Tom, Pfeifer, & Way, 2007), which in turn has the potential to reduce its subsequent influence on behavior. Even when inducing emotion, it is advisable to plan when and how to conduct a manipulation check. For instance, experiences of anger that are labeled as “anger” by participants have a different physiological response pattern than unlabeled experiences of anger (e.g., Kassam & Mendes, 2013). On the other hand, asking participants to retrospectively report their experience later in the experiment also has costs, because memory-based measures have their own biases (Robinson & Clore, 2002).

Images

In daily life, people seek out evocative images in magazines, newspapers, museums, or on the Internet. Researchers use images to induce an affective or specific emotional change in participants (see Table 10.1 for examples). Images from the International Affective Pictures System (IAPS; Lang, Bradley, & Cuthbert, 2008) are most frequently used in psychological research. The major benefit of these images as induction stimuli is that they are normed for affect in both younger (e.g., Ito, Cacioppo, & Lang, 1998; Lang et al., 2008) and older adults (Grühn & Scheibe, 2008), and some images have also been normed for

discrete emotions (Libkuman, Otani, Kern, Viger, & Novak, 2007; Mikels, Frederickson, Larkin, Lindberg, Maglio, & Reuter-Lorenz, 2005). The images have also more recently been normed for distinctiveness, familiarity, and other cognitive/perceptual features (Delplanque, N'diaye, Scherer, & Grandjean, 2007; Libkuman et al., 2007).

In a typical picture induction study, participants are seated in front of a computer screen and shown a series of images, with each one presented for 2–7 seconds followed by an inter-stimulus interval of 50 milliseconds or more. Sometimes participants are shown a class of images in blocks to induce a single, sustained, evocative state (e.g., unpleasant: Lynn, Zhang, & Barrett, 2012). Other times, participants view IAPS images in random order and responses to each image are recorded. Participants can be asked to rate their own experience while viewing the slides (i.e., self-focused emotion), rate the affective or emotional quality of the slides (i.e., world-focused emotion),² or the researcher makes physiological recordings of autonomic nervous system activation and facial muscle movements. Inducing evocative states with visual images is easy and efficient. One major drawback of the IAPS slide set is that they do not sample all portions of affective space equally (there are very few slides to induce low-arousal positive and negative states and high-arousal neutral states). The few IAPS images that appear to be both highly arousing and neutral are only neutral by virtue of their mean ratings across individuals with large standard deviations (meaning that some people experience them as negative and others as positive). A related limitation is that there is considerable idiographic variation across individuals in their affective reactions to the images, although this is not well documented in published research other than by the standard deviations of slide norms. These limitations (uneven distribution of stimuli across the arousal and valence dimensions, and high idiographic variation) are not unique and likely describe affective stimuli more generally. That being said, the IAPS images suffer from a third problem, namely that slides used to evoke pleasant (positive) changes tend to be less arousing than those evoking unpleasant (negative) changes. Finally, IAPS images are also unimodal visual stimuli and do not have the multimodal richness of movies. A recent study found that pairing IAPS images with music was a particularly effective affect induction technique (Lynn et al., 2012), suggesting that future work could combine stimuli to increase the potency of inductions.

Faces

Posed depictions of emotion on the face (scowling faces symbolizing anger, pouting faces symbolizing sadness, etc.) are common in the published literature (see [Table 10.1](#) for references to face sets). Although faces are not routinely used to evoke emotional reactions, they can be used to assess the effect of affect or more specifically emotion on other psychological processes such as visual awareness (e.g., Anderson, Siegel, & Barrett, 2011). In a typical study, participants view digitized images of faces on a computer screen. Some investigators ask participants to watch the faces passively (Lange et al., 2003; block 1), whereas others ask participants to make either emotional judgments of the face (e.g., Critchley et al., 2000, Study 1) or nonemotional judgments (e.g., gender; Critchley et al., 2000, Study 2). It is presumed that either passive viewing or rendering nonemotional judgments involves implicit processing of emotion, whereas labeling a face as emotional brings “online” explicit knowledge about the emotion. Like IAPS images, faces are easy to use in an experiment, but with the main disadvantage that it is not clear which psychological process they provoke. For instance, there is evidence that individuals subtly move their own facial muscles when perceiving another person's facial actions (e.g., Niedenthal, 2007; for a review, see Niedenthal, Mermillod, Maringer, & Hess, 2005), and consistent with this facial feedback hypothesis, some studies find evidence that viewing faces influences self-reported feelings (Dimberg, 1988). Yet it is unclear whether viewing a posed emotional expression induces the same emotion in a perceiver. For example, participants viewing scowling and smiling faces had increased activity in the corrugator supercilii and zygomaticus major facial muscle regions, respectively, but reported experiencing more fear in response to the angry faces and happiness in response to the happy faces (Dimberg, 1988). There are additional concerns with the use of posed, caricatured facial expressions that should make researchers cautious about using them for emotion induction purposes (Barrett, 2011b; Barrett, Mesquita, & Gendron, 2011). The faces might be useful for priming emotion knowledge, however. Posed scowls, pouts, and the like are more like cultural symbols of emotions than inborn, reflexive signals of emotion per se (Barrett, 2011b), suggesting that viewing faces likely activates embodied knowledge about emotion concepts. This claim is supported by evidence that posed facial expressions produce increased activity in brain regions involved in semantic retrieval (Lieberman et al., 2007; Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012). It is also bolstered by event-related potential (ERP) data showing that early potentials distinguish stimuli differing in valence (i.e., pleasantness and unpleasantness), but that later ERPs (i.e., after 300 msec when

semantic information comes “online”) distinguish discrete emotion categories (Eimer & Holmes, 2007).

In some studies faces are followed by a visual mask such that the face is presented very briefly followed by a picture of the same identity posing a neutral face (e.g., Whalen, Rauch, Etcoff, McInerney, Lee, & Jenike, 1998) or a scrambled face (Kim, Loucks, Maital, Davis, Oler, Mazzulla, & Whalen, 2010). Such “backward masking” methods are thought to engage subliminal processing of emotional information, although the mask itself seems to influence how the face stimulus is processed (see Kim et al., 2010). Newer methods for subliminal presentation of faces, such as continuous flash suppression, offer a way of examining the impact of affective changes without the problems associated with backward masking. In continuous flash suppression, two visual images are simultaneously presented via a stereoscope, one image to each eye. One image is static while the other is interleaved in a variety of images that flash and change during brief presentations over the trial. Conscious awareness of the static image is suppressed, and participants only see the flashing images. The unseen image is encoded, however, and has an affective impact that then is misattributed to the affective value of the “seen” image, which is usually objectively neutral (Anderson, Siegel, White, & Barrett, 2012).

Sounds/Voices

Both the acoustical properties of a sound (e.g., its pitch and variation) and its representational meaning (e.g., whether it is the sound of bees, an ambulance siren, or a human voice) evoke affective and emotional changes in perceivers. Specific acoustical properties have the capacity to directly affect the nervous system of the perceiver (Bachorowski & Owren, 2008), whereas other sounds have affective potency because of their conceptual meaning. For example, a gentle buzzing sound might be soothing with prior experiences of bees in a garden, but terrifying if you have been stung by a swarm of bees. The most frequently used sounds for inducing evocative states are the International Affective Digitized Sounds (IADS; Bradley & Lang, 2007) that have been rated in terms of their ability to evoke changes in hedonic valence and level of arousal (see Table 10.1). Participants usually listen to digitized sounds through speakers or headphones. Like other digitized stimuli, sounds are easy to administer. Their major drawback is that, like faces, it is not always clear which psychological processes they provoke (e.g., are they inducing affective changes alone, conceptual changes, or both?). Some require conceptual processing for their

effects and others do not.

There are even more complex issues to consider with vocal stimuli than with other sounds. Although some researchers hypothesize that emotional content is carried by the prosody of a voice (Patel, Scherer, Bjorkner, & Sundberg, 2011) or vocal utterances (Simon-Thomas, Keltner, Sauter, Sinicropi-Yao, & Abramson, 2009), other data suggest that such sounds only carry information about the arousal of the speaker (Bachorowski & Owren, 2008; Russell, Bachorowski, & Fernandez-Dols, 2003). If true, then such stimuli are useful for studying affect rather than emotion per se. In practice, most vocal stimuli are not used to induce an evocative state but are instead used to study emotion perception (i.e., where participants evaluate the emotional or affective meaning of the stimulus). However, models of primate vocal communication suggest that such vocal stimuli shift the affective state of the perceiver (Owren & Rendell, 1997), so it is possible that these stimuli induce a change in affective state in perceivers. The affective or emotional impact of vocal sounds might also vary depending on whether the vocalizations are produced by physiological changes occurring in the speaker, or whether they are volitionally produced even in the absence of changes in affective experience (Owren, Amoss & Rendell, 2011; Scherer, 1995; Scherer, Johnstone & Klasmeyer, 2003), although this distinction remains understudied.

Music

Music is a specific kind of sound used to induce affect and emotion (Juslin & Laukka, 2003; Juslin & Sloboda, 2001). In some studies, music is used alone (e.g., Tamir & Ford, 2009), but in other studies it is often paired with another type of induction stimulus, such as pictures (e.g., Lynn *et al.* 2012) or imagery (e.g., Eich & Metcalfe, 1989). The Continuous Music Technique (CMT; Eich, 1995; Eich & Metcalfe, 1989) is a well-known affect induction (see Table 10.1). A major advantage of the CMT is that music continues to play throughout the experiment, which extends the duration of the evocative state (e.g., Lindquist & Barrett, 2008) and permits the participant to simultaneously perform another task. The major disadvantage of this technique is that it is relatively ineffective for inducing specific emotions, although it can robustly induce affective states. For example, the CMT does not reliably induce distinctive states of anxiety and anger (both unpleasant, highly aroused states), but it reliably induces an unpleasant, highly aroused state, a pleasant state, and a neutral state (see Lench, Flores, & Bench, 2011, table 4).

Imagery and Recall

Imagery and recall are not only used in conjunction with music (e.g., Eich & Metcalfe, 1989), but they can also be used on their own as an effective method for inducing affect and emotion (Lench et al., 2011). Neuroimaging evidence has demonstrated that imagining the future, remembering the past, and creating fictitious imaginings recruit a similar network of brain regions (e.g., Spreng, Mar, & Kim, 2008), suggesting that memory and imagery rely on similar psychological mechanisms that involve retrieval of embodied information from the past. These same brain regions show an increase in activation during the experience of emotion (Kober, Barrett, Joseph, Bliss-Moreau, Lindquist, & Wager, 2008), consistent with our hypothesis that prior experience is important for creating emotional states from simpler affective changes (Barrett, 2006b; Barrett & Bliss-Moreau, 2009b), and indicating that imagery and recall are valid ways to induce affect or emotion. For example, the “scenario induction” technique has been successfully used to evoke a variety of emotional experiences during brain-imaging experiments (Wilson-Mendenhall, Barrett, Simmons, & Barsalou, 2011). Indeed, people frequently engage in “mental time travel” throughout the day, during which they remember emotional events from their past, or imagine emotional events to come in the future, and the resulting affective changes are more potent than those induced by the person's immediate circumstances (Killingsworth & Gilbert, 2010). One of the major benefits of the mental imagery and recall techniques is their relative ease of use. A potential drawback is that participants differ in the ability to engage in mental imagery (e.g., Marks, 1973).

Words

Since Osgood's classic work (e.g., Osgood, Suci, & Tannenbaum, 1957), it has been well known that words have affective connotations, and therefore should have the capacity to produce affective changes in a speaker or a listener. There are standardized sets of evocative words, including the Affective Norms for English Words (ANEW; Bradley & Lang, 1999; see Table 10.1). In a typical experiment, words are presented to participants either supraliminally (i.e., for a second or longer) or subliminally (i.e., latencies under 50 msec) on a computer screen to prime affective content without participants' conscious awareness (Bargh, 2004; also see Bargh and Chartrand, Chapter 13 in this volume). Affective primes have been shown to be generally effective, although affective priming scores have only low to moderate reliability, which can lead to

inconsistency in effects across studies (for review see De Houwer, Tegie-Mocigemba, Spruyt, & Moors, 2009). Also, as noted in the meta-analysis by Lench, Flores, and Bench (2011), priming manipulations have relatively small effect sizes. Furthermore, a method like the sentence-unscrambling task, in which participants reconstruct a set of scrambled words into an emotional sentence, does not, in and of itself, alter emotional state (Innes-Ker & Niedenthal, 2002).

Of all the evocative stimuli used for induction purposes, words are perhaps the easiest to present because they do not require special technology (i.e., even a piece of paper will suffice). It is important to recognize that, like faces and voices, words evoke both changes in representations of the body that are experienced as affective (e.g., Lewis, Critchley, Rotshtein, & Dolan, 2007) but they also require conceptual processes involved in word recognition and comprehension. Consistent with this view, neuroimaging evidence indicates that words are represented as “embodied” – that is, as re-enactments of prior sensory and motor experiences (Kan, Barsalou, Solomon, Minor, & Thompson-Schill, 2003). Despite these findings, meta-analytic evidence suggests that, on average, presenting participants with evocative words can induce anxiety, although this might not be sufficient to induce other evocative states (see Lench et al., 2011, table 4). Like faces, words might be better used as primes to activate conceptual knowledge than as a means to induce feelings per se.

Bodily Movements and Posture

Given the recent emphasis on the role of simulation and embodiment in emotion (e.g., Niedenthal, 2007), it seems reasonable that bodily movements and posture could be used to evoke affect generally and emotion more specifically. For example, the facial feedback hypothesis states that feedback from contraction of specific facial muscles provides affective information to the central nervous system about the affective state being expressed which is then interpreted (see McIntosh, 1996). Early studies utilizing facial movements to induce affect were criticized because of the strong demand characteristics of the task, which meant that participants could have used conceptual knowledge rather than the physical aspects of the task to report an emotion state consistent with the face posed (e.g., Zuckerman, Klorman, Larrance, & Spiegel, 1981). To address this concern, Strack *et al.* (1988) developed a paradigm that believably altered facial muscle activation without invoking conceptual knowledge about emotion by asking participants to hold a pen in their pursed lips, which covertly prevented muscle

activation consistent with a smile, or between their teeth, which activated muscles associated with a smile. For other recent uses of this paradigm, see Supplemental [Table 10.1](#) at <http://www.affective-science.org/publications.shtml>. Beyond moving facial muscles, postural and other gross bodily movements have also been used to induce changes in affective state or, more commonly, to alter affective judgments of stimuli (i.e., world-focused affect). A smaller number of studies have used overall changes in bodily posture (along with careful cover stories to avoid demand characteristics) to directly alter a participant's emotional state (e.g., Stepper & Strack, 1993) and in some cases postures changed brain activity consistent with an approach or avoidance motivational state (e.g., Harmon-Jones & Peterson, 2009). Combining bodily manipulations across multiple body systems (e.g., facial changes with postural changes with imagined or presented evocative stimuli) might further intensify the potency of such manipulations (e.g., Flack, Laird, & Cavallaro, 1999; but see Price & Harmon-Jones, 2010).

Physiological Manipulations

The classic work of Schachter and Singer (1962) demonstrated how pharmacological manipulation of physiological arousal (with injections of epinephrine or placebo) altered the experience of anger versus happiness (depending on a confederate's behavior when the arousal symptoms were unexpected), and changed the participant's behavior (e.g., participant agreed/disagreed with the confederate, or engaged in behaviors initiated by the confederate). Although these findings were interpreted as evidence that social affiliation influenced the construction of anger or happiness specifically, the observed changes are also consistent with a simple manipulation of hedonic valence.³ Other physiological manipulations, such as caffeine, have resulted in weak or no affect-altering effects (and any affective impacts may be attributable to caffeine withdrawal; James & Rogers, 2005). Exercise provides perhaps the most well-characterized way to manipulate peripheral physiological arousal producing an affective change (e.g., Ekkekakis, Parfitt, & Petruzzello, 2011). For example, when exercise intensity reaches the exerciser's ventilatory threshold (i.e., beyond which exercise becomes increasingly anaerobic instead of aerobic), individuals switch from reporting a positive affective state to a negative one (Ekkekakis et al., 2011). Other work has shown that a brief (i.e., 5-minute) bout of cycling exercise alone did not have an affective impact (Tomaka, Blascovich, Kibler, & Ernst, 1997), perhaps because the physiological arousal induced by exercise must be of longer duration to alter subjective experience of

affect, which suggests that endocrine or other bloodborne effects of increased arousal may be critical for a successful induction of this sort.

Although physiological manipulations of affect can be quite potent, they come with the distinct disadvantage that many require considerable expertise to administer and extensive precautions for their safe use, and thus are relatively scarce in the psychological literature. Oxytocin, for example, is administered intranasally in humans, and has recently emerged as a potential way to manipulate affect. It has been shown to decrease arousal ratings of visual images of human, but not animal, threat stimuli (Norman, Cacioppo et al., 2011). Much of the research to date has investigated the effects of oxytocin on the perception of affect and emotion (e.g., Gamer, Zurowskis, & Buchels, 2010) rather than on emotion induction, but the work by Norman *et al.* (2011) suggests it may be a promising affect inducer or modulator, although perhaps only in the presence of social stimuli (for a review, see Norman, Hawkey, Cole, Berntson, & Cacioppo, 2011).

Botulinum neurotoxin-A (i.e., botox), used cosmetically to reduce facial wrinkles, is a peripheral physiological method for changing affect. Most commonly, botox injections into the corrugator supercilii muscle region (i.e., “scowl” muscles) have been used to alter affective ratings of evocative videos (Davis, Senghas, Brandt, & Ochsner, 2010) and decrease depression (Finzi & Wasserman, 2006). Botox injections to the corrugator region also decreased activation in the left amygdala when individuals imitated scowling facial expressions, and more generally decreased coupling between the amygdala and dorsal brain stem areas responsible for autonomic efferent activity (Hennenlotter, Dresel, Castrop, Ceballow-Baumann, Wohlschlager, & Haslinger, 2009).

Emerging methodologies for directly manipulating brain activity are expanding the potential to manipulate affect via the central nervous system. For example, “real time functional magnetic resonance imaging” (rtfMRI) allows researchers to detect (with fMRI) and provide feedback to a person about their ongoing brain activity as they experience a mental state (e.g., Weiskopf, Veit, Erb, Mathiak, Grodd, Goebel, & Birbaumer, 2003; Yoo & Jolesz, 2002). With feedback, participants gain the ability to regulate activity in brain regions associated with affect such as the insula (e.g., Caria, Veit, Sitaram, Lotze, Weiskopf, Grodd, & Birbaumer, 2007) or related areas such as the anterior cingulate cortex to modulate the experience of pain (DeCharms et al., 2005). Also, the future will likely bring greater use of transcranial magnetic stimulation

(TMS) in which a magnetic pulse is used to temporarily activate or disrupt activity in certain brain areas. Here, researchers measure experiential or behavioral changes when a brain area is temporarily stimulated or taken “offline.” For instance, a study used TMS of the anterior temporal lobe that is thought to support semantic judgments, among other things, to show that participants were significantly slower to complete a task that required them to find a matching synonym in a set of words than in a control task of similar difficulty (Lambon Ralph, Pobric, & Jefferies, 2009). To date, TMS has been used to study the perception of facial expressions (Pitcher, Garrido, Walsh, & Duchaine, 2008), motor cortex excitability during affective picture viewing (Hajcak, Molnar, George, Bolger, Koola, & Nahas, 2007), and approach-avoidance tendencies (Schutter, de Weijer, Meuwese, Morgan, & van Honk, 2008).

Confederates

Schachter and Singer (1962) published arguably the most famous emotion study to utilize confederates, but labs have been using scripted confederates to induce emotion or affect for the past several decades (see Table 10.1; Cohen, Nisbett, Bowdle, & Schwarz, 1996; DeSteno, Bartlett, Baumann, Williams, & Dickens, 2010). Confederates typically produce impactful changes in induced affect and emotion. Designs using confederates are labor-intensive, however, involving lots of practice to ensure that confederates are convincing and that their behavior is the same across participants. In addition, researchers must attend to such details as controlling the confederate's vocal prosody and nonverbal behaviors, and carefully scripting the confederate's behavior and words. Use of additional lab equipment (videotape or microphone) helps a researcher ensure that every administration is as similar as possible.

Motivated Performance Tasks

In a motivated performance task, participants give an impromptu speech in front of an audience (Trier Social Stress Test or TSST; Kirschbaum, Pirke, & Hellhammer, 1993) or complete serial subtraction problems in the presence of an evaluative experimenter (e.g., Quigley, Barrett, & Weinstein, 2002) to produce high arousal affective states and alter autonomic nervous system activity. (See Table 10.1 for variations on these methods.) The advantage of motivated performance tasks is that they are ecologically valid and both subjectively and physiologically evocative. The robust nature of motivated performance tasks can

also be a disadvantage because certain participants (particularly those with social anxiety or low self-esteem) might find them excessively distressing and may even disengage from the task altogether. Researchers must therefore take precautions at screening and also use methods for detecting when an individual has disengaged and is no longer performing the task.

Virtual Reality

In virtual reality, participants (or players) are presented with digital (and sometimes photorealistic) images of what look like real-world people, objects, scenes, and events, which are combined with tracking of the player's movements to allow her or him to become immersed in and interact with this artificial world as if it were real. Virtual reality allows a person to immerse themselves in a social situation or a scene in a first-person way (as opposed to viewing the scene in a third-person way) – a distinction that appears to have specific neural correlates (e.g., Ochsner, Knierim, Ludlow, Hanelin, Ramachandran, Glover, & Mackey, 2004; Ruby & Decety, 2004). Although virtual reality has great potential as an affect and emotion induction method, this method is, thus far, used rarely. A notable exception is Project EMMA (Engaging Media for Mental Health Applications) in Spain that examines how emotion contributes to “presence” (feeling part of, or immersed in) of a virtual environment. Here, a virtual urban park with multisensory features (e.g., sounds, sights, different kinds of affective stimuli) is used to induce changes such as anxious, relaxed, or neutral affective states (Riva et al., 2007). For details and a use of these methods for another application, psychotherapy, see Table 10.1. Computer-based virtual reality, other immersive technologies like augmented reality in which photorealistic objects are combined with computer-simulated environments and/or objects, and other related technologies, like gaming, are likely to radically change affect and emotion research (for an excellent and accessible look at this revolution, see Blascovich & Bailenson, 2011). Blascovich *et al.* (2002) enumerated the methodological advantages of virtual-reality-based studies for social psychological research including research in emotion and affect (see Table 10.1). These methods are likely to provide a potent way to induce affect or emotion because the human brain is wired to “travel” to virtual worlds (using the “default” network) in the form of remembering the past, imagining the future, and mind wandering beyond one's current circumstances (Andrews-Hanna, Reidler, Huang, & Buckner, 2010). In some ways, imagination is a low tech type of “virtual reality” that appears to rely on the same brain circuitry.

Real-World Stimuli

Researchers have used spiders, snakes, participation in extreme sports, foods or other substances, pain stimuli, and odors or other chemosensory stimuli to induce affect and emotional changes (see [Table 10.1](#) for methods). Experience-sampling methods (also known as diary methods, ecological momentary assessment, or ambulatory assessment) are useful for tracking these real-world objects and events that have the capacity to induce affective and emotional changes. Details on experience-sampling methods, supporting technology, and analysis methods for the interested reader can be found in Mehl and Conner ([2012](#)) and in Reis, Gable, and Maniaci, [Chapter 15](#) in this volume.

Measuring Evoked States

Measuring general affective and more specific emotional changes is complex and fraught with difficulties. A persistent challenge is that many researchers implicitly use the measurement model depicted in [Figures 10.1a](#) or [10.1b](#) (called an effect-indicator model), which is consistent with classical measurement theory (cf., Barrett, [2000](#); Barrett, [2006a](#), [2011a](#)). In this view, a stimulus triggers a latent emotional state indexed by a set of measured variables that are strongly correlated with one another (because of their common cause). In such a model, an emotion, such as anger, would have a characteristic facial expression (e.g., a scowl), a characteristic body change (e.g., an increase in heart rate with an increase in blood pressure), and a characteristic change in subjective experience (e.g., fury), and each of these measures would be strongly correlated with one another (because of their common, latent cause). Each emotion category is assumed, in essence, to be a psychological “type” with a biological core. If emotions worked this way, then it would only be necessary to measure one observable (e.g., facial muscle movements, cardiovascular changes, or self-reports of experience) because the others would be redundant with it (being so highly correlated). Although the evidence is strongly suggestive that measurements of valence taken across different measurement modalities do correlate with one another, as do different measures of arousal, and that positive affect seems to have a distinct profile from negative affect, the same cannot be said for anger, sadness, fear, disgust, or happiness as discrete emotional states (Barrett, [2006a](#); Cacioppo, Berntson, Larsen, Poehlmann, & Ito, [2000](#); Lindquist et al., [2012](#); Mauss & Robinson, [2009](#)). Given the tremendous variation in instances within an emotion category (such that sometimes blood pressure goes

up, sometimes it does not; sometimes a person approaches, at other times they withdraw), it is necessary to capture and model individual emotional instances.

a) Example of an effect indicator model: A basic emotion model

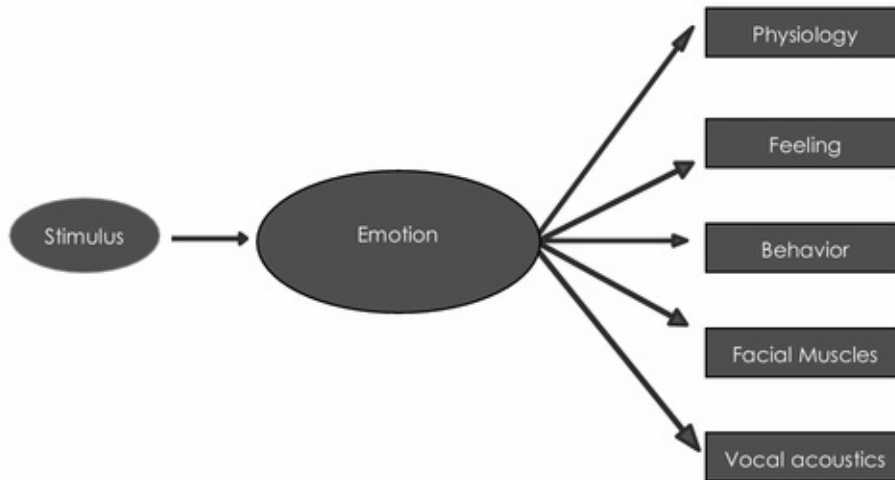


Figure 10.1a. Basic emotion models (see Gendron & Barrett, 2009 for review) assume that an emotion (e.g., fear) is a reflex triggered by an external stimulus (e.g., a bear) that results in a coordinated array of outputs that unfold over time.

b) Example of an effect indicator model: An appraisal emotion model

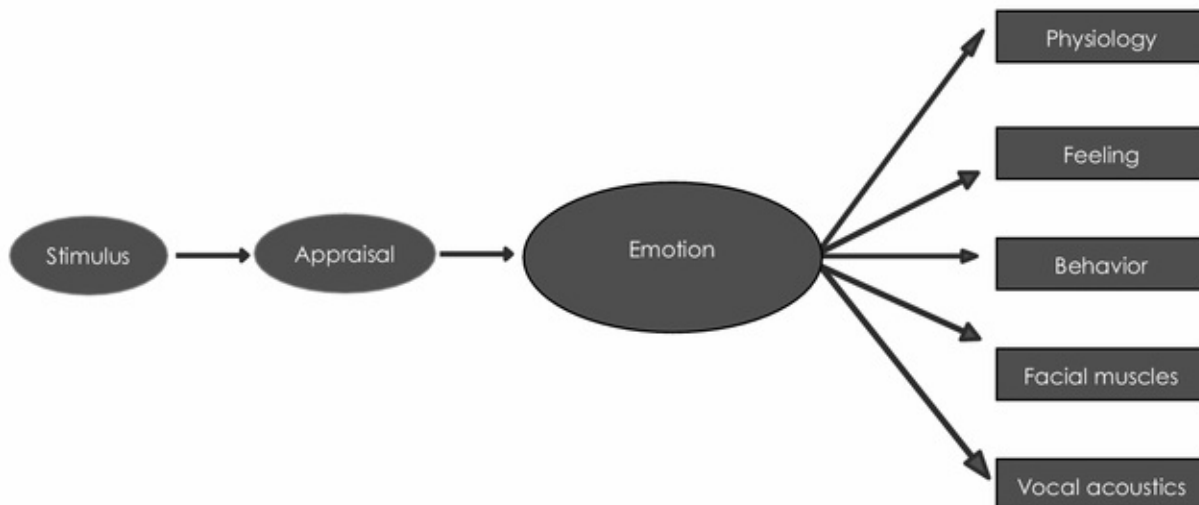


Figure 10.1b. Some types of appraisal models (i.e., causal variants) assume that a cognitive appraisal of the stimulus (e.g., a bear is threatening) results in a specific emotional response (e.g., fear) that results in a coordinated array of outputs that unfold over time (see Gendron & Barrett, 2009, for review). In this sense, causal appraisal models are like basic emotion models (see 1a). Other types of appraisal models (i.e., constitutive variants) assume that an appraisal is a description of experience during emotional episodes and are more like

psychological constructionist models (see 1c).

c) A causal indicator model of emotion

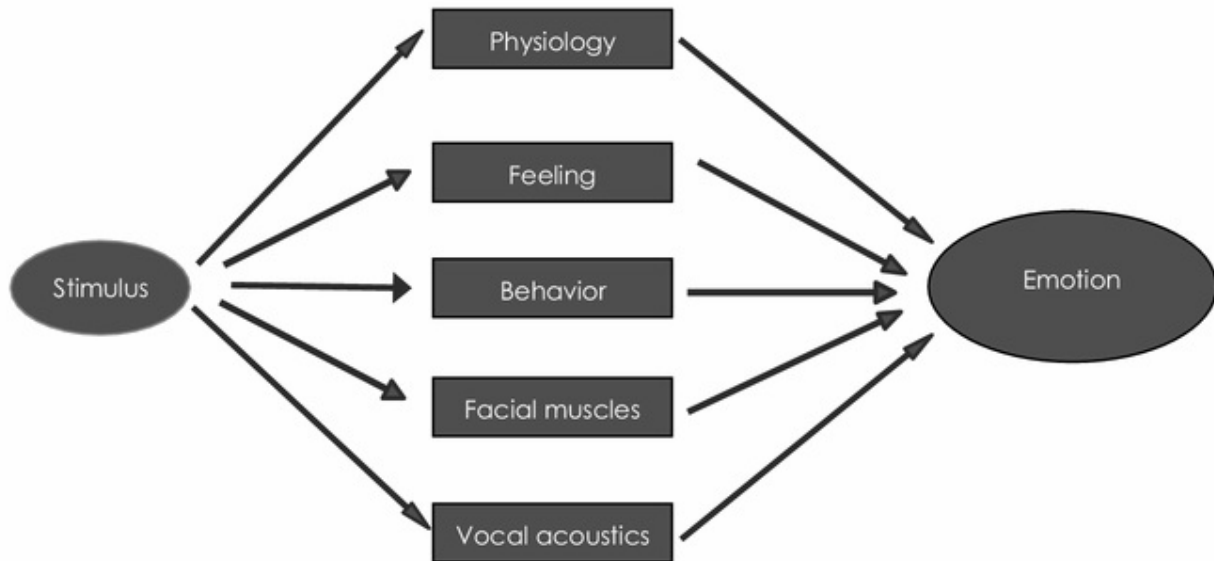


Figure 10.1c. A psychological constructionist model (e.g., Barrett, 2011a; Russell, 2003) posits that the "stimulus" is comprised of core affect, sensory input from the world and conceptual knowledge supported by language that work together to produce an emergent state that is an emotion (or more correctly an emotional instance). Thus, in the model, indicators are not assumed to correlate, but rather together they instantiate the current emotion state.

d) Example of a causal indicator model: A psychological construction model of emotion

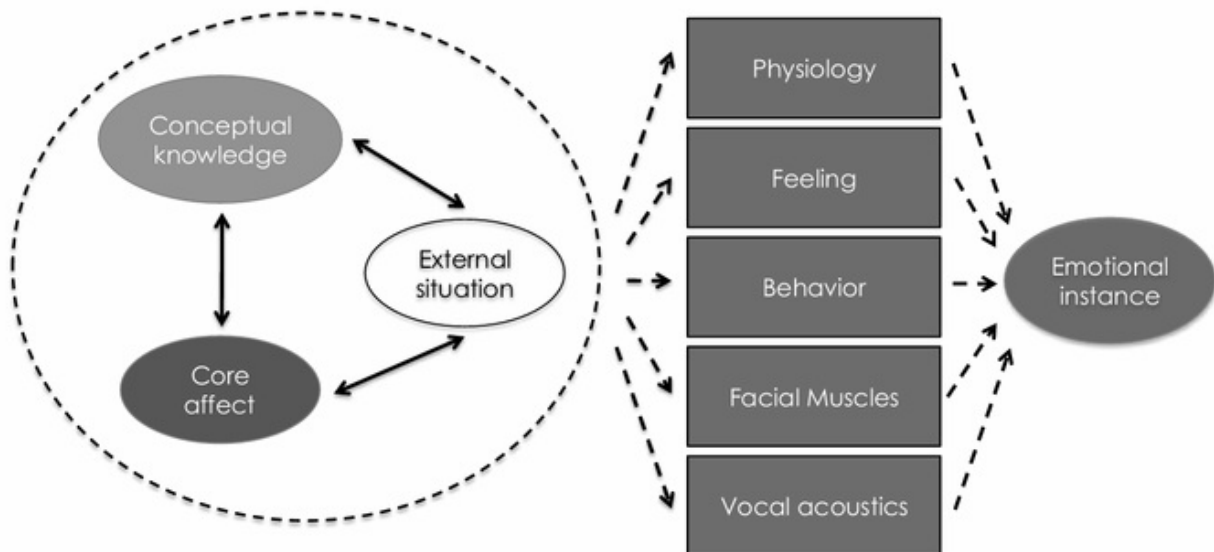


Figure 10.1d. To illustrate details of the formal model in 1c., our psychological constructionist model (Barrett, 2006a, b; 2009; 2011a; Barrett,

Lindquist & Gendron, 2007; Barrett, *et al.* 2007; Lindquist et al., 2012) specifically posits that core affect from the body (red), sensory input from the world (white) and conceptual knowledge supported by language (blue) work together to produce an emergent state that can be measured as a discrete emotion (purple). In a given instance of emotion (e.g., *anger at a spouse*), the constellation of measures will take one pattern and in another instance (e.g., *anger at an injustice*), it will take a different pattern.

An alternative to the effect indicator model of emotion is to measure multiple modalities within a single study and combine them using a causal indicator model (see Figure 10.1c for the formal model and Figure 10.1d for an exemplar model; as explained by Barrett 2000, 2011a; Coan 2010). In this measurement approach, measures are not expected to correlate with one another, but instead their aggregate realizes or constitutes an instance of the latent construct in question (for a discussion of latent constructs using the “causal indicator” approach, and how these latent constructs differ theoretically from those estimated with the more popular and familiar “effect indicator modeling” approach, see Barrett, 2011a; Bollen & Lennox, 1991). By definition, in the causal indicator approach, instances of emotion within the same emotion category can vary from one another without violating the assumptions of the latent construct model. Furthermore, an instance of emotion can only be measured using more than one measurement modality, and using only one measure (e.g., skin conductance), which constitutes a violation of the measurement model. This approach (which is usually applied to modeling socioeconomic status, for example) is well suited to the study of emotion where subjective reports, physiological measurements, and behavioral observations rarely, if ever, strongly correlate (Barrett, 2006b). For measures, references, and the major advantages and disadvantages of each set of measures, see Table 10.2 and a more extensive Supplemental Table 10.2 at <http://www.affective-science.org/publications.shtml>.

Facial Muscle Activity

Facial electromyography (or facial EMG) measures facial muscle activity that varies as a function of whether someone is in a pleasant or an unpleasant state. Interestingly, because the skin serves as a low pass filter of the muscle activations occurring beneath the skin's surface, very small changes in facial EMG can be detected that do not necessarily result in externally observable

movement of the features of the face (e.g., Cacioppo, Bush, & Tassinary, 1992; Cacioppo, Petty, Losch, & Kim, 1986; Tassinary & Cacioppo, 1992). Thus, facial EMG provides a tool for detecting very subtle facial muscle activation even if the participant later inhibits or otherwise aborts the full expression of an initiated facial response. A meta-analysis by Cacioppo *et al.* (2000) showed that facial EMG can frequently though not invariantly distinguish pleasant from unpleasant affective states (e.g., Cacioppo, Martzke, Petty, & Tassinary, 1988). Unpleasant affective states are most likely to be associated with increased activation over the corrugator supercilii muscle region (e.g., Schwartz, Fair, Salt, Mandel, & Klerman, 1976), whereas pleasant affective states are most likely to be associated with activation over the zygomaticus major muscle region (e.g., Harmon-Jones & Allen, 2001). However, there are no consistent and specific facial EMG-based “signatures” for specific emotion states such as anger, fear, or disgust (for reviews, see Barrett, 2011b; Russell et al., 2003), despite the fact that posed expressions are used in emotion perception research.

Table 10.2. Measures, References, and Advantages and Disadvantages of Methods for Measuring the Impact of Affect and Emotion Inductions

Measurement Domain	Measures (with typical abbreviations) and References	Advantages	Disadvantages
Facial Muscle Activity	<ul style="list-style-type: none"> For methodological details on recording facial EMG, see e.g., (Fridlund & Cacioppo, 1986; Tassinari, Cacioppo, & Geen, 1989; Tassinari, Cacioppo, & Vanman, 2007). Facial electromyography (fEMG) to measure emotion or affective variables includes such sites as the zygomaticus major muscle region ("smiling") and the corrugator supercilii muscle region ("scowling") 	<ul style="list-style-type: none"> Sensitive to subtle and/or fleeting changes in muscle activation. The ability to distinguish positive from negative affective states. 	<ul style="list-style-type: none"> Special equipment and expertise are needed Fairly elaborate cover stories are needed to prevent participants from guessing the true nature of the experimental questions since muscle activity is voluntarily controlled Skin preparation can be tedious and uncomfortable for participants We can currently only measure with reasonable specificity a small number of muscle regions in the face Participants typically make more facial muscle movements when another person is present (or implied (Fridlund, 1991)) There are no consistent and specific facial EMG-based "signatures" for specific emotion states such as anger, sadness, fear, or disgust
Vocal Acoustics	<ul style="list-style-type: none"> For an in-depth discussion of measuring vocal acoustics, see Owren & Bachorowski (2007) 	<ul style="list-style-type: none"> Provides an observer-independent measure 	<ul style="list-style-type: none"> Requires some specialized equipment and expertise
Observer Ratings	<ul style="list-style-type: none"> The most popular coding system is the Facial Action Coding system (FACS; Ekman & Friesen, 1978). Observers rate activity in each of 44 facial action units (AUs). The AUs are visible facial muscle movements and are singly, or in combinations, hypothesized to be characteristic of specific emotions. The Maximal Descriptive Facial Movement Coding System (MAX) for use with infants (Izard, 1979) Child Facial Coding System for coding facial pain expressions (Gilbert et al., 1999) The Facial Expression Coding System (FACES) for facial muscle movements related to affect rather than discrete emotions; facial actions are rated as positive/negative and for intensity (Kring & Sloan, 1991). 	<ul style="list-style-type: none"> Uses a standardized measurement tool 	<ul style="list-style-type: none"> Typically time-consuming and resource-intensive

Measurement Domain	Measures (with typical abbreviations) and References	Advantages	Disadvantages
Subjective Experience	<ul style="list-style-type: none"> • Example measures to assess affect include an affect grid (Russell et al., 1989), rating dial, or joystick to measure each of the dimensions of affective state or Self-Report Manikins (Bradley & Lang, 1994). • Example measures to assess emotion include the Current Mood Questionnaire (Barrett & Russell, 1998), the Positive Affect and Negative Affect Scale-Extended (Watson & Clark, 1994), and the Differential Emotions Scale (DES; Izard, Dougherty, Bloxom, & Kotsch, 1974) 	<ul style="list-style-type: none"> • Self-report is currently the only valid way of assessing subjective experience 	<ul style="list-style-type: none"> • Measures of discrete emotional states tend to measure pleasant or dysphoric affect (although there are notable individual differences)

Measurement Domain	Measures (with typical abbreviations) and References	Advantages	Disadvantages
Behavioral Changes	<ul style="list-style-type: none"> Behaviors used to measure affect include: approach (e.g., push a lever toward the stimulus) or avoid tendencies (pull a lever away from a stimulus) as an index of positive or negative feelings toward that stimulus (e.g., Chen & Bargh, 1999), or whether someone consumed a drink as evidence of positive affect (Winkielman, Berridge, & Wilbarger, 2005). Behaviors used to index the experience of specific emotions include pride measured as a greater tendency to persevere on difficult tasks (Williams & DeSteno, 2008), fear measured as greater risk aversion (Lerner & Keltner, 2001; Lindquist & Barrett, 2008a), or social behaviors, e.g., cooperation during the experience of gratitude (DeSteno et al., 2010) or jealousy (DeSteno et al., 2006). 	<ul style="list-style-type: none"> The potential to measure “unconscious” emotion or affect Easily measurable behaviors for affect as approach or avoidance states Behaviors are often observer-independent measures 	<ul style="list-style-type: none"> No one-to-one correspondence between behaviors and discrete emotional states
Autonomic Nervous System Activity	<ul style="list-style-type: none"> For an introduction, see Stern, Ray, and Quigley (2001). For more advanced users, see Cacioppo, Tassinary, and Berntson (2007). Papers on specific systems/measures are available at: http://www.sprweb.org/journal/index.cfm#guidelines Common measures include heart rate (HR; inverse of heart period), blood pressure (BP), cardiac output (CO), total peripheral resistance (TPR), stroke volume (SV), respiratory sinus arrhythmia (RSA, also known as high frequency heart period variability), electrodermal activity (EDA; including event-related skin conductance responses [ERA-SCRs], nonspecific skin conductance responses [NS-SCRs], or tonic skin conductance level [SCL]), respiratory rate, tidal volume (V_T), the electrogastrogram (EGG), pupillary diameter, or face or hand temperature. 	<ul style="list-style-type: none"> Measures can distinguish positive from negative affective states. The measures are often observer-independent, in that most of them are not at all, or only minimally, affected by volitional changes on the part of the participant 	<ul style="list-style-type: none"> There are no consistent and specific patterns of autonomic response for specific emotion states such as anger, fear, sadness, or disgust. These measures require equipment and expertise. These measures are resource intensive both in preparing participants for recordings, and in the reduction of data post-acquisition. Measures require careful thought regarding the nature of the psychological state that can be inferred from the physiological measures.

Measurement Domain	Measures (with typical abbreviations) and References	Advantages	Disadvantages
Central Nervous System Changes	<ul style="list-style-type: none"> • Methods include electroencephalography (EEG) from which one can derive event-related potentials (ERPs), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), positron emission tomography (PET), and more recently, functional near-infrared spectroscopy (fNIRS). For methods, see Fabiani, Gratton, and Federmeier (2007); Pizzagalli, (2007). • Both EEG and MEG result from electrical activity in the brain that is the net effect of ionic currents flowing between neurons across the synapse. • EEG is a measure of electrical changes in the brain recorded as voltage changes and MEG is a measure of magnetic field changes at the scalp • Event-related electrical or magnetic changes to affective or emotional stimuli are event-related potentials (ERPs) in EEG studies or event-related magnetic fields (ERFs) in MEG studies • Source imaging can be used with MEG (MEG and MRI paired) or EEG to better localize measures to a specific anatomical structure 	<ul style="list-style-type: none"> • These measures reveal something about the processes underlying affect and emotion that are not necessarily accessible via conscious self-report or observable behavior (for an example, see Lindquist et al., 2012). 	<ul style="list-style-type: none"> • Require “reverse inference” • Expense and access to equipment • Complex data analysis • Extensive need for expertise in data acquisition, data analysis and neuroanatomy • Cannot achieve optimal temporal and spatial resolution simultaneously • Emotions cannot be clearly and unambiguously assessed (i.e., measures do not reliably differentiate anger from sadness from fear; Lindquist et al., in press). • Concerns with false-positive findings attributable to the typical use of multiple comparisons across voxels in the brain
Endocrine, Immune, and Inflammatory Changes	<ul style="list-style-type: none"> • Example measures: anger and testosterone; (Peterson & Harmon-Jones, 2011); immunoglobulin A (immune factor in saliva), and disgust (Stevenson, Hodgson, Oaten, Barouei, & Case, 2011); basal levels of the pro-inflammatory cytokine, IL-18 and negative affect with a sadness induction (Prossin et al., 2011); IL-6 response to a motivated performance task eliciting anger and anxiety (Carroll et al., 2011) 	<ul style="list-style-type: none"> • Provides a peripheral physiological measure that goes beyond what can be measured using traditional psychophysiological measures 	<ul style="list-style-type: none"> • Some measures are difficult to obtain in the typical psychological lab • Requires control over numerous extraneous variables, e.g., factors like time of day, time of last meal, menstrual cycle phase, etc. • Assays can be expensive • The temporal characteristics of measures are slow relative to the brief nature of affective and emotional changes

Vocal Acoustics

Vocal acoustics (i.e., the auditory parameters of a person's speech) are sometimes used to index a person's affective state, particularly the sender's level of arousal (for a review, see Bachorowski & Owren, 2008). Although some researchers argue that certain patterns of vocal acoustics correspond consistently and specifically to certain emotional states (e.g., Patel et al., 2011), other

summaries of the literature refute that claim (e.g., Russell et al., 2003). Even studies claiming that specific vocal acoustics differentiate emotions tend to find evidence for more basic underlying dimensions that characterize the vocal acoustics across emotions. Patel *et al.* (2011) recently found three dimensions corresponding to the physiological processes involved in the production of vocal sounds (e.g., one dimension characterized by pressure on the subglottis and vocal fold adduction, one by the quality of vocal fold adduction, and one by either low or high mean frequency of the vocal output). At least one of these dimensions (subglottal pressure/vocal fold adduction) seems related to arousal because it distinguishes sounds made during the experience of relief from sounds made during joy, anger, and fear (Patel et al., 2011). Measures of vocal acoustics provide an observer-independent assessment of affective or emotional state. As with several of these measures, however, assessing vocal acoustics requires specialized equipment and expertise (see Table 10.2).

Observer Ratings

Researchers often attempt to measure emotion in the laboratory by asking trained or untrained raters to infer a participant's mental state by observing his or her behavior. Implicit in asking a perceiver to make such judgments is the assumption that each emotion has a prototypical expression displayed in the face, voice, or body for all the world to see (i.e., it is assumed that faces, voices, and body movements are “read-outs” or “signals” of an emotional state). The majority of studies reporting that non-expert perceivers are able to “recognize” emotional behaviors typically have experimental methods that include contextual constraints that lead to a higher percentage of judgments that agree with the experimenter's expectations (such as providing a limited number of emotion words and having perceivers choose the relevant term from this smaller set; for evidence on the importance of emotion words in producing accurate emotion perceptions, see Barrett, Lindquist, & Gendron, 2007; Barrett et al., 2011; Gendron, Lindquist, Barsalou, & Barrett, in press; Lindquist & Gendron, 2013). Often perceivers are asked to distinguish two emotions that differ in valence (e.g., anger vs. happiness) or arousal (e.g., anger vs. sadness), such that affective distinctions are actually driving the observed effects and it cannot be concluded that emotion differences are present. Of note, facial expressions usually occur only when another person is present (e.g., Fernández-Dols & Ruiz-Belda, 1995; Russell et al., 2003) or in the implied presence of another person (Fridlund, 1991). This suggests that facial expressions are more like communicative symbols than signals of specific mental states (see Barrett, 2011b).

Behavior

When using behaviors to index the internal state of a participant, it is important to remember that doing so is essentially a formalized instance of theory of mind. Just as all human perceivers infer intentionality, desires, goals, and personality traits to other humans by observing their behavior (Malle & Holbrook, 2012), experimenters infer these mental states in their participants. In experiments that aim to measure emotion, experimenters typically rely on prototypical scripts to link behaviors to mental states, with the underlying assumption that a given behavior indicates the presence of a single emotion category. This assumption is hard to justify in mammals, which have considerable behavioral flexibility and tremendous behavioral variability within any emotion category (e.g., aggression or withdrawal could indicate fear). For example, rats do many things in threatening or dangerous situations that could correspond with fear; they freeze (e.g., LeDoux, Iwata, Cicchetti, & Reis, 1988), startle (e.g., Hitchcock & Davis, 1987), avoid the threat (e.g., Vazdarjanova & McGaugh, 1998), or attack (e.g., Blanchard, Hori, Rodgers, Hendrie, & Blanchard, 1989), and each of these so-called fear behaviors is produced by a distinct neural circuit and has distinct autonomic nervous system correlates that prepare the body for action. The specific behavior emitted fits the immediate situation with which the animal must cope. Similarly, when measuring emotional behavior in humans, we need to consider a priori which behavior will best allow the participant to cope with the constraints of the experimental situation, which may or may not be the same as the “prototypic” emotional behavior prescribed by the script. The same holds true for measuring affect – if the situation demands it, people can approach even when threatened (Jamieson, Koslov, Nock & Mendes, 2013).

Autonomic Nervous System Activity

For more than a century, scientists have attempted to use psychophysiological measures to assess affect and emotion. These measures (e.g., changes in heart rate or blood flow) are often controlled by both the sympathetic nervous system, which when activated, often results in greater arousal, and the parasympathetic nervous system, which when activated, often results in reduced arousal. Most scientists agree that autonomic changes are integral to affect and emotion. Yet it is important to realize that many non-affective or nonemotional states (e.g., involving attention, mental effort, etc.) also result in autonomic changes.³ In fact, both branches of the autonomic nervous system are involved in energy management (i.e., the sympathetic nervous system, when activated, results in

greater catabolic activity or greater use of energy stores, and the parasympathetic nervous system, when activated, results in greater anabolic or energy-conserving processes). Similarly, cortisol, often considered a “stress” hormone in the psychological literature, is important for managing metabolic activity in the body. This observation implies that changes in affect and emotion have direct implications for energy balance and maintaining homeostasis.

Certain “myths” about the autonomic nervous system prevail in the emotion and affect literature and have led to misperceptions, methodological problems, and unwarranted inferences when interpreting results (see [Table 10.3](#)). Perhaps the most important misconception is that discrete emotions like anger, sadness, fear, disgust, and happiness can be distinguished by consistent and specific autonomic signatures. Cacioppo *et al.* (2000) provided a thorough meta-analysis of the then-extant literature on the psychophysiology of emotion, which, along with other recent reviews (e.g., Barrett, 2006b; Lindquist *et al.*, in press), suggested that there are no consistent and specific autonomic signatures for discrete emotions, although autonomic measures can sometimes distinguish a person in a positive versus negative state (Cacioppo *et al.*, 2000), a threat versus a challenge state (Quigley *et al.*, 2002; Tomaka, Blascovich, Kibler, & Ernst, 1997), or whether someone is highly aroused or not (Bradley, Codispoti, Cuthbert, & Lang, 2001). Other summaries of the literature note that it is important to consider situational context when interpreting the emotional meaning of autonomic changes (Kreibig, 2010).

Table 10.3. Common Myths Observed in Studies of Emotion and Affect that Measure Autonomic Nervous System Activity

Myth 1. Autonomic nervous system arousal, particularly in the sympathetic nervous system, is a unitary construct.

One of the most pervasive assumptions about the autonomic nervous system is that arousal is unitary, leading some to assume that a single measure of function or activation will suffice to represent autonomic arousal across the entire body. This cannot be assumed. This arose from early physiological work (e.g., Cannon, 1915, 1932), suggesting that activation in the sympathetic branch of the autonomic nervous system was predominant under conditions of bodily activation, and that it exerted highly coordinated action on organs throughout the body. Instead, it is now clear in humans and nonhuman animals, in

particular among mammals, that there is target-specific and exquisitely tuned control of changes in activation of both the sympathetic and parasympathetic nervous systems. Although a more generalized activation of sympathetic outflows can occur, this typically happens under intensely evocative circumstances. A nice demonstration of the regional specificity of sympathetic activation was shown in a study in which investigators used microneurography (i.e., peripheral nerve recordings in awake humans) to record muscle sympathetic nerve activity simultaneously in a participant's leg and arm. In this study, mental arithmetic increased activation of the sympathetic nerves to muscles in the leg, but did not simultaneously alter sympathetic nerve activity to the arm (Anderson, Wallin, & Mark, 1987). For a useful review of the regional and organ specificity of sympathetic nervous system activity, see Morrison (2001).

Myth 2. Sympathetic activation is always accompanied by parasympathetic withdrawal (or vice versa).

This myth is another legacy of Cannon's writings. We now know that not all activation in the sympathetic and parasympathetic nervous systems is reciprocally coupled (i.e., a pattern of increased activity in one autonomic branch accompanied by decreased activity or withdrawal, in the other branch [for discussion, see Berntson et al., 1991]). Although reciprocal coupling is common, it is not ubiquitous. Nonreciprocal modes of control can occur as an increase or decrease in activity in one autonomic branch with no change in activity of the other branch, or even as simultaneous activation or inhibition of both autonomic branches. Coactivation has been demonstrated in both humans and rats during attentional orienting (Gianaros & Quigley, 2001; Quigley & Berntson, 1990). Several authors have suggested that coactivation and coinhibition likely have important functional consequences (Berntson et al., 1991; Paton, Boscan, Pickering, & Nalivaiko, 2005).

Myth 3. Changes in skin conductance specifically reflect changes in arousal.

Few measures of autonomic function have been as popular for measuring emotional or affective states as skin conductance (or more

broadly, electrodermal activity). For example, Lang and colleagues have consistently shown that the magnitude of skin conductance responses to International Affective Picture Set (IAPS) slides and other stimuli is related to changes in the self-reported arousal elicited by these stimuli (e.g., Bradley et al., 2001). Although the eccrine sweat glands have the advantage of receiving input from only the sympathetic branch of the autonomic nervous system and correlating positively with self-reported arousal, skin conductance also is responsive to numerous physical conditions including temperature, humidity and skin hydration, and to many mental states including the relative familiarity vs. novelty of a stimulus, mental effort, *etc.* To permit strong inferences about the psychological process of interest, experimenters using skin conductance measures must carefully control contextual and stimulus variables (Cacioppo & Tassinary, 1990).

Myth 4. Affective or emotional states are accompanied only by efferent outflow from the brain to the peripheral, autonomically innervated target organs, without impact on afferent inputs to the brain.

Psychophysiological autonomic measures are often interpreted as if they only reflect efferent autonomic outflow from the central nervous system to the periphery. However, affective autonomic responses result from the delicate interplay between afferent and efferent nerve traffic over time. Measures of organ function will reflect (within seconds) both efferent outflow from the central nervous system and afferent inflow to the central nervous system from organs like the heart and gastrointestinal tract. Unfortunately, our understanding of afferent (or interoceptive) impacts and our ability to measure them, especially in humans, is less well developed than our ability to measure peripheral target organ changes. This makes it difficult to distinguish co-occurring efferent and afferent effects. Fortunately, brain imaging studies can now provide at least some composite information about afferent peripheral activation during affective states (e.g., Critchley, 2005).

Myth 5. Autonomic changes in the body only exist to support affective or emotional states.

This is, of course, an overstatement. It is not uncommon, however, for researchers to fail to consider that physiological measures must be interpreted in view of the overall, concurrent functioning of the body. Autonomic functions subserve not just our affective states but our very survival. This does not mean that affective states are not themselves critical to survival, but rather that they occur in the context of other basic functions like breathing, movement of blood through the body and digestion of food, all of which happen concurrent with our changing affective and emotional states.

When using psychophysiological measures, researchers should carefully consider the epoch over which the affective or emotional response is measured. Autonomic responses in the laboratory typically will have the largest amplitude when an affective event is initiated, and often (but not always) amplitudes diminish as the stimulus continues. Because autonomic changes are the predominant means by which the body produces the initial, fast changes in a peripheral organ like the heart or lungs (i.e., on the order of milliseconds to seconds), autonomic effects will predominate over these shorter time periods. Slower-acting physiological systems (e.g., endocrine or immune changes) will predominate when stimuli are extended (e.g., minutes to hours). Physiological systems also have a dynamic range (i.e., minimum to maximum) under normal physiological conditions. If the basal state of autonomic activation is near one end of the physiological range, there can be physiological constraints on reactivity, which must be considered. For example, if an individual's basal heart rate is near either end of the dynamic range of one of the autonomic branches, as might occur for heart rate when a person is standing (i.e., where basal sympathetic activity is high, see Berntson, Cacioppo, & Quigley, 1993), then an affect induction may not be able to cause much further sympathetically mediated increase in heart rate (for a discussion, see Berntson, Cacioppo, & Quigley, 1991). It is also important to eliminate or statistically control for substances participants may have ingested that can impact their autonomic responses to affective stimuli. Examples include medications and non-medicinal substances like alcohol, caffeine, or illicit drugs. In addition, researchers should screen for chronic diseases and acute illnesses that could impact autonomic function either directly (e.g., diabetes or heart disease) or because medications commonly used to treat these diseases have autonomic effects (e.g., asthma). Even in young, healthy participants, these precautions will reduce variability and enhance the

researcher's ability to detect affectively induced autonomic changes, which is critical given the notoriously high variability of physiological measures.

Central Nervous System Activity

Affect can be measured by recording electrical, metabolic, or hemodynamic changes in the brain and researchers consistently attempt to use these measures to measure emotion. The use of these methods in the science of affect and emotion is hotly debated because they rely on “reverse inference,” or the idea that it is possible to infer a mental state from the measurement of a physical state (see note in [Table 10.3](#), Myth 3 concerning the same issue when making psychophysiological inferences; see also Cacioppo & Tassinary, 1990). Different measures (e.g., electroencephalography [EEG], event-related potentials [ERPs], magnetoencephalography [MEG], functional magnetic resonance imaging [fMRI], and positron emission tomography [PET]) provide somewhat different information about brain activity, and there are common misperceptions about what can be inferred about affect and emotion with these methods. Because both electrical and magnetic measurements of changes under the scalp's surface (EEG/ERP and MEG, respectively) have some spatial imprecision, they can only localize the source of signals to larger brain areas (i.e., relative to fMRI, which is better at localizing activation to more specific coordinates in space); MEG has slightly better spatial resolution than EEG/ERP, because magnetic fields are less distorted by the skull and scalp than are electrical fields (Cohen, Nisbett, Bowdle, & Schwarz, 1990; Leahy, Mosher, Spencer, Huang, & Lewine, 1998). Poor spatial specificity makes it hard to localize the signal to specific brain structures or spatial locations in the brain (which, when known, can be useful for understanding what psychological processes might be invoked during a given experiment). Although it has limited spatial resolution, the temporal resolution of EEG and MEG is on the order of milliseconds. Thus, EEG/ERP and MEG are ideal for revealing the time course of affective and emotional events, but less suited for spatial localization than fMRI or PET. The hardware costs and physical space constraints are fewer for EEG, so it has benefits over MEG in this regard.

Compared with studies using fMRI or PET, relatively fewer studies have used EEG/ERP to investigate changes in affective or emotional experiences, and even fewer have used MEG (although see, e.g., Morel, Ponz, Mercier, Vuilleumier, & George, 2009). Perhaps the best-known series of studies to use EEG to investigate emotion have assessed the lateralization of responses to pleasant and

unpleasant affect. These studies generally link pleasant affect to relatively greater electrical activity in the left frontal lobe and unpleasant affect to relatively greater activity in the right frontal lobe (Ahern & Schwartz, 1985; Davidson, Ekman, Saron, Senulis, & Friesen, 1990). Studies have also assessed the lateralization of anger experience (e.g., Harmon-Jones & Allen, 1998, 2001). More commonly, researchers use ERP methods to study emotion perception (as participants are viewing posed, caricatured facial expressions, e.g., a scowl for anger, a pout for sadness, etc.). The evidence from these studies suggests that early ERPs (80–180ms) reflect the categorization of a face as a face (vs. non-face), as generally affective (neutral vs. valenced), as positively versus negatively valenced, or as displaying some degree of arousal (Eimer & Holmes, 2007; Palermo & Rhodes, 2007). Other studies find that later components (peak activations up until 230 msec) are differentially sensitive to anger and fear faces that are incongruously paired with fear and anger body postures (Meeren, Van Heijnsbergen, & DeGelder, 2005). These findings suggest that these later components reflect a distinction between discrete emotions, because a person would have to perceive that faces were depicting anger versus fear in order to experience the face and body postures as incongruous in this task. Of interest, the time window required for distinguishing among different discrete emotions is approximately the same as that required for semantic processing of other visual stimuli (e.g., Schmitt, Münte, & Kutas, 2000).

As discussed by Berkman, Cunningham, and Lieberman (Chapter 7 in this volume), fMRI measures hemodynamic activity in the brain (i.e., blood flow inferred from changes in blood oxygen levels), and PET is a measure of metabolic changes (i.e., most commonly, glucose metabolism), which can be assessed during affective or emotional tasks. Relative to MEG or EEG, fMRI and PET have poorer temporal resolution because there is a lag of several seconds between stimulus onset and resulting hemodynamic or metabolic changes (e.g., the hemodynamic response reflects not only blood flow changes to a given stimulus, but also the influences of whatever occurred for about 30 seconds beforehand). However, fMRI and PET have better spatial resolution, and are thus better for studies concerned with the spatial location of neural activation during evocative events. A growing number of studies have investigated the brain basis of affect and emotion predominantly using fMRI. Emerging meta-analytic evidence indicates that positive and negative affect show different patterns of neural activity, although different meta-analyses do not consistently agree on what those differences are (Kringelbach & Rolls, 2004; Wager et al., 2008). Analyses generally agree, however, that discrete

emotional states such as anger, sadness, fear, disgust, and happiness do not show consistent and specific increases in neural response during the experience of discrete emotions (Lindquist et al., 2012, although see Vytal & Hamann, 2010, for a different perspective). Instead, fMRI/PET data support the idea that there are a set of more fundamental psychological building blocks that, in combination, give rise to the variety of discrete emotional states (Barrett, Mesquita, Ochsner, & Gross, 2007; Kober et al., 2008; Lindquist et al., 2012).

Endocrine, Immune, and Inflammatory Changes

A growing number of studies use changes in endocrine, immune, or inflammatory markers in an attempt to measure affect or emotion. Endocrine, immune, and inflammatory measures provide a broader assessment of peripheral physiological change and can be obtained alongside more traditional autonomic nervous system measures. They do, however, have the distinct disadvantages of being expensive and potentially difficult to obtain in the typical psychology lab, require control over multiple extraneous variables (at minimum, statistically, for factors like time of day or when the person last ate), and require considering how to minimize the possibility that taking a sample itself will induce an affective change (e.g., pain from a needle stick or disgust induced by providing a saliva sample). Endocrine and immune system changes occur on the order of minutes to hours, making their temporal features less optimal for detecting the typically fast (i.e., milliseconds to seconds) and frequently more fleeting changes evoked by affective or emotional stimuli.

Subjective Experiences

In principle, it should be possible to use objective measures of emotion (in the face, body, or brain) to measure how a person is feeling without asking for a self-report. If emotions should be measured and modeled using an “effect indicator” latent model as depicted in Figure 10.1a, then aspects of an emotional response are connected by a single common cause, and it should be possible to measure the more easily observable aspects of emotion (e.g., facial movements, vocal acoustics, peripheral physiology) to learn something about a person's subjective experience (which itself is not observable without a self-report). Furthermore, using an effect indicator model, when there is lack of correspondence between verbal reports and these objective measurements (as there almost always is), researchers often assume that the verbal reports are invalid. Similarly, if a person says he is angry but pouts (which is typically

perceived as sadness), researchers usually would believe him to feel sad, because behavior would trump verbal report as a way of indexing subjective experience. In practice, objective measures in the brain and body tend to be weakly correlated with one another, and together they do not consistently and specifically distinguish between instances of anger, sadness, fear, and so on (Barrett, 2006a; Barrett, Lindquist et al., 2007; Lindquist et al., 2012). As a result, objective measures cannot be used as proxy measures of emotional experience. Scientists are not able to use any single measurement, or profile of measurements, to indicate when a person is feeling anger, fear, sadness, or anything similar. If we want to know whether a person is experiencing an emotion, we have to ask her/him. Verbal reports are inappropriate for revealing the *processes* that produce subjective experiences (i.e., how emotions are caused), but barring social desirability concerns, they are the only way to assess the *content* of subjective experiences of emotion (i.e., what people are feeling; Barrett, 2006b; Barrett, Mesquita et al., 2007).

When asking a participant to characterize subjective experiences, most researchers simply present a set of adjectives and ask the participant to rate how well each word describes his or her immediate feeling state (for a list of typical measures and references, see Table 10.2). This rating process assumes that the feeling state is static and can be held constant while it is compared to different emotion or affective concepts to produce the best match, so that the process of comprehending and rating emotion or affect-related words will not change the experience at hand. It is possible, even likely, however, that thinking about emotion adjectives can change how a participant feels, rather than just reflect that feeling, and so adjective rating scales should be used judiciously. Furthermore, what appears to be a simple judgment actually draws on a set of complex processes including (1) the participants' access to phenomenal or "raw" experience, (2) his or her ability to verbalize this experience as "reflective" feelings that can be communicated in awareness, (3) knowledge of the emotion words and related emotion concepts represented by the words, (4) having sufficient executive attention resources to move from item to item to render a set of ratings, and (5) social desirability concerns.

With these points in mind, there are important considerations when using adjective scales to measure subjective experience. First, participants will report how they are feeling using whatever measure a researcher gives them, regardless of what the scale is called, even when the items are not entirely appropriate. For example, if a participant is feeling angry, but is given the Beck Depression Inventory (Beck & Steer, 1987), she will likely use the items given to

communicate how unpleasant they feel. Thus, it is important to measure both the emotion of interest and other closely related emotions for discriminant validity. Second, there are individual differences in emotional granularity, or the extent to which people represent their experiences in distinctive categorical terms. Minimally, this means that not everyone is able to report on the difference between a sad, angry, guilty, or any other feeling, but it also suggests that some people don't *feel* these experiences distinctly and instead experience more general affective changes (Barrett, 1998, 2004; Barrett & Bliss-Moreau, 2009b; Feldman, 1995). As a result, some individuals use emotion words to refer to distinct experiences, whereas others use the same words to represent their feelings in more basic affective terms (that is, they use the same words for what those words have in common, which is unpleasant feeling).

In addition to asking people to describe their emotional experiences with a set of emotion words, it is also possible to assess emotional experiences by measuring how people judge the world around them during an emotional episode. Sometimes these are called emotional “appraisals” (e.g., Akinola & Mendes, 2008; Lerner & Keltner, 2001), but this is also a case of “world-focused” emotion (Lambie & Marcel, 2002; Lindquist & Barrett, 2008). In the appraisal approach to emotion, appraisals are often thought of as the cognitive mechanism that automatically evaluates a stimulus, which in turn triggers a specific emotion (Ellsworth & Scherer, 2003). But from another theoretical perspective, appraisal judgments reflect world-focused experiences of emotion by describing how a person experiences the world during a particular emotional episode (cf., Barrett, Mesquita et al., 2007; for a consistent theoretical view, see Clore & Ortony, 2008). For instance, during fear, people (at least in a Western cultural context) experience a world full of risk (e.g., Lerner & Keltner, 2001; Lindquist & Barrett, 2008b). In anger, they experience others as blameworthy.

There continues to be debate regarding whether or not a person can feel both pleasant and unpleasant at the same time (for a discussion, see Barrett & Bliss-Moreau, 2009a), with no resolution of this debate in sight. Therefore, an experimenter has to make an explicit decision as to whether hedonic valence will be measured with one bipolar item (ranging from pleasant to unpleasant) or two unipolar items (ranging from neutral to pleasant and neutral to unpleasant). It is important to keep in mind that many participants impose bipolarity on ambiguously unipolar scales – for example, how sad you are, anchored from “not at all” to “intensely,” where “not at all” is interpreted by many respondents as “happy” (Carroll & Russell, 1996). This problem is reduced, but not eliminated, by explicitly labeling scale anchors. Further, although there

continues to be debate over the theoretically most valid way to parse affective space (e.g., Cacioppo & Gardner, 1999; Russell & Barrett, 1999), all affective properties (valence/arousal, approach/avoid, positive activation/negative activation) are related to one another and can be derived from one another (Carroll, Yik, Russell, & Barrett, 1999; Yik, Russell, & Barrett, 1999) as long as the entire affective space is adequately sampled (Barrett & Russell, 1998).

Finally, the issue of response scaling goes well beyond the debates about bipolarity. Concerns about how people use Likert-type scales are gaining momentum in the science of self-report (e.g., Bartoshuk, 2000; Bartoshuk, Fast & Snyder, 2005), and so scale considerations should be carefully considered in any study that involves the measurement of subjective experience. Many studies simply have participants indicate the extent to which an adjective describes his or her immediate feeling state on a scale from low to high (e.g., 1 = not at all, 5 = very much). Recent work by Bartoshuk *et al.* (2005) indicates that there are strong individual differences in how people interpret such anchors and use such scales, going well beyond the old discussions of response styles. As a result, some researchers are now adopting a general labeled magnitude scale approach, where vague Likert-type scale choices are explicitly anchored to an absolute set of comparisons, to allow different individuals to be calibrated to one another in their scale usage (Bartoshuk, 2000).

Tips, Tricks, and Secrets for “Best Practices”

A psychologist's task is to discover facts about the mind (e.g., changes in affect or emotion) by measuring responses from a person (e.g., reaction times, perceptions, eye or muscle movements, bodily changes, or perhaps electrical, magnetic, blood flow, or chemical measures related to neurons firing). In so doing, psychologists use ideas (in the form of concepts, categories, and constructs) to transform their measurements into something meaningful. The relation between any set of numbers (reflecting a property of the person, or the activation in a set of neurons, a circuit, or a network) and a psychological construct depends on a set of theoretical assumptions. All scientists make such assumptions, whether or not they explicitly express them. First and foremost, then, it is critical for researchers to be clear and explicit about their guiding theoretical framework. Theory not only prescribes a strategy for analysis and interpretation, but it also guides what stimuli can be used for an induction, the dependent variables to be measured, as well as when and how manipulation checks are to be performed. Having an explicit theoretical view of emotion also

maximizes the possibility that the researcher will make design choices that permit strong inferences about the psychological processes at work as reflected in the measures observed. Researchers also must be attentive to the methodological limitations of their chosen induction and measurement methods; the goal may be to induce and measure an emotional state, but the findings might only permit inferences about affect.

Let us consider briefly two different examples, one in which the scientific question is about affect more generally, and the other in which the question concerns a specific emotional state, such as “fear,” to make explicit some of the considerations needed when designing an affect vs. an emotion study. If negative affect is the phenomenon of interest, then, as we noted earlier, it will be especially critical to ensure that activation of conceptual knowledge about specific emotions is minimal or nil so as to permit making inferences solely about negative affect without the confound that the participant activated a particular emotional concept like “fear.” A focus on negative affect also requires the researcher to be cautious about the timing and nature of manipulation checks so that conceptual knowledge about particular emotions is not activated too early and thereby impact the affect induction. If instead we are interested in studying the impact of the specific emotion state of “fear,” then we must also consider how and when a stimulus primes or activates conceptual knowledge about that emotion state. We also need to consider the possibility that even within an emotion category like fear, there can be tremendous variation in the objective responses measured across individuals as a function of individual variability or the context within which fear is elicited. We submit that this is not a bug due to the experimental design, but rather a feature of how the emotion system is built such that different responses are evoked when circumstances call for different adaptations required for meeting particular goals. Also, when studying a specific emotion like fear, then the researcher must also induce and compare appropriate “control” emotions that differ from the focal emotion on dimensions of valence (e.g., by inducing anger or another negative emotion). These emotions experimentally control for the possibility that any effects that appear to be stemming from fear are not simply a function of just any negatively valenced state. And note that researchers will be on the firmest inferential grounds for interpreting their measured face, voice, bodily, or central nervous system outcomes by not just inducing two negatively valenced emotions, but also by equating them for the induced arousal. Lastly, to make claims about a response being specific to a given emotion (e.g., fear), researchers should rule out the possibility of having evoked another emotion with the same valence (e.g.,

anger); in other words, the fear induction should specifically induce fear and not anger, and the reverse should be true for the anger induction.

Finally, when using biological measures to try and index general affective states, or more specific emotional states, it is important to remember that peripheral physiology was not engineered to help us express emotion – it evolved for homeostasis and metabolic regulation. This means that only a small proportion of the variance in biological measures reflects changes in mental states. Furthermore, bodily state measures such as measures of heart rate or skin conductance have their own limitations. These include often being multiply determined by both sympathetic and parasympathetic autonomic changes (i.e., heart rate) that make the autonomic determinants unclear, being sensitive to many psychological effects other than just affect or emotion (e.g., familiarity of stimuli, prior learning about stimuli), or even just being affected by changes in the physical environment (e.g., skin conductance can be altered by the humidity and temperature of the testing room). The limitations and caveats of each induction type and measurement modality must be considered in making inferences and in ruling out potential confounding effects. In sum, following these suggested guidelines and utilizing the resources summarized here based on our current state of knowledge should lead us toward a more valid and replicable science of affect and emotion.

¹ Even the distinction between “cognition” and “emotion” is culturally relative (e.g., Lutz, 1985; for a discussion, see Barrett, 2009).

² Researchers often describe images (or other stimuli such as music, odors, other people, etc.) as “beautiful,” or “distasteful,” with the assumption that pleasure or displeasure is an inherent quality of the stimulus. Stimuli are only pleasant, or distasteful, however, because they alter a perceiver's affect in some way (Barrett & Bliss-Moreau, 2009a). Nonetheless, people often experience affect as a literal property of a stimulus, and we can ask participants to report on the affective or emotional qualities of a stimulus (i.e., world-focused; Lindquist & Barrett, 2008) or on their own state (i.e., self-focused). The caveat about manipulation checks noted in the film section applies to images as well – labeling the emotional content of an unpleasant picture during viewing reduces subsequent self-reported distress to that picture (Lieberman, Inagaki, Tabibnia, & Crockett, 2011), so researchers should consider carefully when and how to

measure subjective responses.

³ In addition, a so-called third branch of the autonomic nervous system, the enteric nervous system, is a specialized nerve plexus lying with the walls of the gastrointestinal system that controls motility and secretion in parts of the intestinal tract and receives modulatory input from the two primary autonomic nervous system branches (Grundy & Schemann, 2007). Activity of this branch is rarely measured in studies of emotion or affect, although it represents a potential novel avenue for future research.

References

- Ahern, G. L., & Schwartz, G. E. (1985). Differential lateralization for positive and negative emotion in the human brain: EEG spectral analysis. *Neuropsychologia*, 23(6), 745–755.
- Akinola, M., & Mendes, W. B. (2008). The dark side of creativity: Biological vulnerability and negative emotions lead to greater artistic creativity. *Personality and Social Psychology Bulletin*, 34(12), 1677–1686.
- Allen, R., & Smith, M. (1997). *Film theory and philosophy*. New York: Oxford University Press.
- Anderson, E., Siegel, E. H., & Barrett, L. F. (2011). What you feel influences what you see: The role of affective feelings in resolving binocular rivalry. *Journal of Experimental Social Psychology*, 47, 856–860.
- Anderson, E., Siegel, E. H., White, D., & Barrett, L. F. (2012). Out of sight but not out of mind: Unseen affective faces influence evaluations and social impressions. *Emotion*, 12(6), 1210–1221.
- Anderson, E. A., Wallin, B. G., & Mark, A. L. (1987). Dissociation of sympathetic nerve activity in arm and leg muscle during mental stress. *Hypertension*, 9 (Suppl III), III114–III119.
- Andrews-Hanna, J. R., Reidler, J. S., Huang, C., & Buckner, R. L. (2010). Evidence for the default network's role in spontaneous cognition. *Journal of Neurophysiology*, 104(1), 322–335.
- Bachorowski, J. A., & Owren, M. J. (2008). Vocal expressions of emotion. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions*

- (3rd ed., pp. 196–210). New York: Guilford Press.
- Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3), 614–636.
- Bargh, J. A. (2004). The four horsemen of automaticity: Awareness, attention, efficiency, and control in social cognition. In J. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (2nd ed., pp. 1–40). Hillsdale, NJ: Lawrence Erlbaum.
- Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition and Emotion*, 12(4), 579–599.
- Barrett, L. F. (2000). *Modeling emotion as an emergent phenomenon: A causal indicator analysis*. Paper presented at the Annual Meeting of the Society for Personality and Social Psychology, Nashville, TN.
- Barrett, L. F. (2004). Feelings or words? Understanding the content in self-report ratings of experienced emotion. *Journal of Personality and Social Psychology*, 87(2), 266–281.
- Barrett, L. F. (2006a). Are emotions natural kinds? *Perspectives on Psychological Science*, 1(1), 28–58.
- Barrett, L. F. (2006b). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, 10(1), 20–46.
- Barrett, L. F. (2009). The future of psychology: Connecting mind to brain. *Perspectives on Psychological Science*, 4(4), 326–339.
- Barrett, L. F. (2011a). Bridging token identity theory and supervenience theory through psychological construction. *Psychological Inquiry*, 22(2), 115–127.
- Barrett, L. F. (2011b). Was Darwin wrong about emotional expressions? *Current Directions in Psychological Science*, 20(6), 400–406.
- Barrett, L. F., & Bliss-Moreau, E. (2009a). Affect as a psychological primitive. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 41, pp. 167–218). Burlington: Academic Press.
- Barrett, L. F., & Bliss-Moreau, E. (2009b). Variety is the spice of life: A psychological construction approach to understanding variability in emotion. *Cognition and Emotion*, 23(7), 1284–1306.

- Barrett, L. F., Lindquist, K. A., Bliss-Moreau, E., Duncan, S., Gendron, M., Mize, J., & Brennan, L. (2007). Of mice and men: Natural kinds of emotions in the mammalian brain? A response to Panksepp and Izard. *Perspectives in Psychological Science*, 2(3), 297–311.
- Barrett, L. F., Lindquist, K. A., & Gendron, M. (2007). Language as context for the perception of emotion. *Trends in Cognitive Sciences*, 11(8), 327–332.
- Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, 20, 286–290.
- Barrett, L. F., Mesquita, B., Ochsner, K. N., & Gross, J. J. (2007). The experience of emotion. *Annual Review of Psychology*, 58, 373–403.
- Barrett, L. F., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology*, 74, 967–984.
- Bartoshuk, L. M. (2000). Comparing sensory experiences across individuals: Recent psychophysical advances illuminate genetic variation in taste perception. *Chemical Senses*, 25, 447–460.
- Bartoshuk, L. M., Fast, K., & Snyder, D. J. (2005). Differences in our sensory worlds: Invalid comparisons with labeled scales. *Current Directions in Psychological Science*, 14(3), 122–125.
- Beaupré, M. G., & Hess, U. (2005). Cross-cultural emotion recognition among Canadian ethnic groups. *Journal of Cross-Cultural Psychology*, 36(3), 355–370.
- Beck, A. T., & Steer, R. A. (1987). *BDI, Beck Depression Inventory: Manual*. San Antonio, TX: Psychological Corp.
- Berntson, G. G., Cacioppo, J. T., & Quigley, K. S. (1991). Autonomic determinism: The modes of autonomic control, the doctrine of autonomic space, and the laws of autonomic constraint. *Psychological Review*, 98(4), 459–487.
- Berntson, G. G., Cacioppo, J. T., & Quigley, K. S. (1993). Cardiac psychophysiology and autonomic space in humans: Empirical perspectives and conceptual implications. *Psychological Bulletin*, 114(2), 296–322.
- Blanchard, C. D., Hori, K., Rodgers, J. R., Hendrie, C. A., & Blanchard, R. J.

- (1989). Attenuation of defensive threat and attack in wild rats (*Rattus rattus*) by benzodiazepines. *Psychopharmacology*, 97, 392–401.
- Blascovich, J., & Bailenson, J. (2011). *Infinite reality: Avatars, eternal life, new worlds, and the dawn of the virtual revolution*. New York: Harper Collins.
- Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, 13(2), 103–124.
- Bliss-Moreau, E., Owren, M. J., & Barrett, L. F. (2010). I like the sound of your voice: Affective learning about vocal signals. *Journal of Experimental Social Psychology*, 46(3), 557–563.
- Bollen, K. & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110(2), 305–314.
- Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation I: Defensive and appetitive reactions in picture processing. *Emotion*, 1(3), 276–298.
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), 49–59.
- Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. University of Florida: The Center for Research in Psychophysiology.
- Bradley, M. M., & Lang, P. J. (2007). *The International Affective Digitized Sounds (2nd Edition; IADS-2): Affective ratings of sounds and instruction manual*. University of Florida.
- Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., & Ito, T. A. (2000). The psychophysiology of emotion. In R. Lewis & J. M. Haviland-Jones (Eds.), *The handbook of emotion* (2nd ed., pp. 173–191). New York: Guilford Press.
- Cacioppo, J. T., Bush, L. K., & Tassinary, L. G. (1992). Microexpressive facial actions as a function of affective stimuli: Replication and extension. *Personality and Social Psychology Bulletin*, 18(5), 515–526.
- Cacioppo, J. T., & Gardner, W. L. (1999). Emotion. *Annual Review of*

Psychology, 50, 191–214.

- Cacioppo, J. T., Martzke, J. S., Petty, R. E., & Tassinary, L. G. (1988). Specific forms of facial EMG response index emotions during an interview: From Darwin to the continuous flow hypothesis of affect-laden information processing. *Journal of Personality and Social Psychology*, 54(4), 592–604.
- Cacioppo, J. T., Petty, R. E., Losch, M. E., & Kim, H. S. (1986). Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of Personality and Social Psychology*, 50(2), 260–268.
- Cacioppo, J. T., & Tassinary, L. G. (1990). Inferring psychological significance from physiological signals. *American Psychologist*, 45(1), 16–28.
- Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2007). *The handbook of psychophysiology* (3rd ed.). New York: Cambridge University Press.
- Cannon, W. B. (1915). *Bodily changes in pain, hunger, fear and rage*. New York: Appleton.
- Cannon, W. B. (1932). *The wisdom of the body*. New York: Norton.
- Caria, A., Veit, R., Sitaram, R., Lotze, M., Weiskopf, N., Grodd, W., & Birbaumer, N. (2007). Regulation of anterior insular cortex activity using real-time fMRI. *NeuroImage*, 35, 1238–1246.
- Carroll, J. E., Low, C. A., Prather, A. A., Cohen, S., Fury, J. M., Ross, D. C., & Marsland, A. L. (2011). Negative affective responses to a speech task predict changes in interleukin (IL)-6. *Brain, Behavior, and Immunity*, 25(2), 230–231.
- Carroll, J. M., & Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology*, 70(2), 205–218.
- Carroll, J. M., Yik, M. S. M., Russell, J. A., & Barrett, L. F. (1999). In the psychometric principles of affect. *Review of General Psychology*, 3(1), 14–22.
- Carter, L. E., McNeil, D. W., Vowles, K. E., Sorrell, J. T., Turk, C. L., Ries, B. J., & Hopko, D. R. (2002). Effects of emotion on pain reports, tolerance and physiology. *Pain Research & Management*, 7(1), 21–30.
- Castanier, C., LeScanff, C., & Woodman, T. (2011). Mountaineering as affect regulation: The moderating role of self-regulation strategies. *Anxiety, Stress,*

and Coping, 24(1), 75–81.

- Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin*, 25(2), 215–224.
- Clore, G. L., & Ortony, A. (2008). Appraisal theories: How cognition shapes affect into emotions. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (pp. 618–644). New York: Guilford Press.
- Coan, J. A. (2010). Emergent ghosts in the emotion machine. *Emotion Review*, 2(3), 274–285.
- Coan, J. A., & Allen, J. J. B. (Eds.). (2007). *Handbook of emotion elicitation and assessment* (Vol. 13). New York: Oxford University Press.
- Cohen, D., Cuffin, B. N., Yunokuchi, K., Maniewski, R., Purcell, C. *et al.* (1990). MEG vs. EEG localization test using implanted sources in the human brain. *Annals of Neurology*, 28(6), 811–817.
- Cohen, D., Nisbett, R. E., Bowdle, B. F., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honor: An “experimental ethnography”. *Journal of Personality and Social Psychology*, 70(5), 945–960.
- Critchley, H. D. (2005). Neural mechanisms of autonomic, affective, and cognitive integration. *The Journal of Comparative Neurology*, 493, 154–166.
- Critchley, H. D., Daly, E. M., Bullmore, E. T., Williams, S. C. R., Amelsvoort, T. V. *et al.* (2000). The functional neuroanatomy of social behaviour: Changes in cerebral blood flow when people with autistic disorder process facial expressions. *Brain*, 123, 2203–2212.
- Davidson, R. J., Ekman, P., Saron, C. D., Senulis, J. A., & Friesen, W. V. (1990). Approach-withdrawal and cerebral asymmetry: Emotional expression and brain physiology: I. *Journal of Personality and Social Psychology*, 58(2), 330–341.
- Davidson, R. J., Scherer, K. R., & Goldsmith, H. H. (2003). *Handbook of affective sciences*. New York: Oxford University Press.
- Davis, J. I., Senghas, A., Brandt, F., & Ochsner, K. N. (2010). The effects of Botox injections on emotional experience. *Emotion*, 10(3), 433–440.
- De Houwer, J., Tegie-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit

- measures: A normative analysis and review. *Psychological Bulletin*, 135(3), 347–368.
- DeCharms, R. C., Maeda, F., Glover, G. H., Ludlow, D. H., Pauly, J. M. *et al.* (2005). Control over brain activation and pain learned by using real-time functional MRI. *Proceedings of the National Academy of Sciences*, 102(51), 18626–18631.
- Delplanque, S., N'diaye, K., Scherer, K., & Grandjean, D. (2007). Spatial frequencies or emotional effects? A systematic measure of spatial frequencies for IAPS pictures by a discrete wavelet analysis. *Journal of Neuroscience Methods*, 165(1), 144–150.
- DeSteno, D., Bartlett, M., Baumann, J., Williams, L. M., & Dickens, L. (2010). Gratitude as moral sentiments: Emotion-guided cooperation in economic exchange. *Emotion*, 10, 289–293.
- DeSteno, D., Valdesolo, P., & Bartlett, M. Y. (2006). Jealousy and the threatened self: Getting to the heart of the green-eyed monster. *Journal of Personality and Social Psychology*, 91(4), 626–641.
- Dimberg, U. (1988). Facial electromyography and the experience of emotion. *Journal of Psychophysiology*, 2(4), 277–289.
- Duclos, S. E., Laird, J. D., Schneider, E., Sexter, M., Stern, L., & Van Lighten, O. (1989). Emotion-specific effects of facial expressions and postures on emotional experience. *Journal of Personality and Social Psychology*, 57(1), 100–108.
- Eich, E. (1995). Searching for mood dependent memory. *Psychological Science*, 6(2), 67–75.
- Eich, E., & Metcalfe, J. (1989). Mood dependent memory for internal versus external events. *Journal of Experimental Psychology*, 15(3), 443–455.
- Eimer, M., & Holmes, A. (2007). Event-related brain potential correlates of emotional face processing. *Neuropsychologia*, 45(1), 15–31.
- Ekkekakis, P., Parfitt, G., & Petruzzello, S. J. (2011). The pleasure and displeasure people feel when they exercise at different intensities: Decennial update and progress towards a tripartite rationale for exercise intensity prescription. *Sports Medicine*, 41(8), 641–671.

- Ekman, P., & Cordano, D. (2011). What is meant by calling emotions basic. *Emotion Review*, 3(4), 364–370.
- Ekman, P., & Friesen, W. V. (1978). *The facial action coding system (FACS): A technique for measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.
- Ellsworth, P. C., & Scherer, K. R. (2003). Appraisal processes in emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 572–595). New York: Oxford University Press.
- Fabiani, M., Gratton, G., & Federmeier, K. D. (2007). Event-related brain potentials: Methods, theory, and applications. In J. T. Cacioppo, L. G. Tassinary & G. G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed., pp. 85–119). New York: Cambridge University Press.
- Feldman, L. A. (1995). Valence focus and arousal focus: Individual difference in the structure of affective experiences. *Journal of Personality and Social Psychology*, 69(1), 153–166.
- Ferguson, M. J., Bargh, J. A., & Nayak, D. A. (2005). After-affects: How automatic evaluations influence the interpretation of subsequent, unrelated stimuli. *Journal of Experimental Social Psychology*, 41(2), 182–191.
- Fernández-Dols, J. M., & Ruiz-Belda, M. A. (1995). Are smiles a sign of happiness? Gold medal winners at the Olympic Games. *Journal of Personality and Social Psychology*, 69(6), 1113–1119.
- Finzi, E., & Wasserman, E. (2006). Treatment of depression with botulinum toxin A: A case series. *Dermatologic Surgery*, 32, 645–650.
- Flack, W. F., Laird, J. D., & Cavallaro, L. A. (1999). Separate and combined effects of facial expressions and bodily postures on emotional feelings. *European Journal of Social Psychology*, 29(2–3), 203–217.
- Fridlund, A. J. (1991). The sociality of solitary smiles: Effects of an implicit audience. *Journal of Personality and Social Psychology*, 60, 229–240.
- Fridlund, A. J., & Cacioppo, J. T. (1986). Guidelines for human electromyographic research. *Psychophysiology*, 23(5), 567–589.
- Gamer, M., Zurowskis, B., & Buchels, C. (2010). Different amygdala subregions mediate valence-related and attentional effects of oxytocin in humans.

Proceedings of the National Academy of Sciences, 107(20), 9400–9405.

Gendron, M., & Barrett, L. F. (2009). Reconstructing the past: A century of ideas about emotion in psychology. *Emotion Review*, 1, 1–24.

Gendron, M., Lindquist, K. A., Barsalou, L. W., & Barrett, L. F. (2012). Emotion words shape emotion percepts. *Emotion*, 12(2), 314–325.

Gianaros, P. J., & Quigley, K. S. (2001). Autonomic origins of a nonsignal stimulus-elicited bradycardia and its habituation in humans. *Psychophysiology*, 38, 540–547.

Gilbert, C. A., Lilley, C. M. C., McGrath, P. J., Court, C. A., Bennett, S. M., & Montgomery, C. J. (1999). Postoperative pain expression in preschool children: Validation of the Child Facial Coding System. *Clinical Journal of Pain*, 15(3), 192–200.

Gross, J. J., & Levenson, R. W. (1995). Emotion elicitation using films. *Cognition and Emotion*, 9(1), 87–108.

Grühn, D., & Scheibe, S. (2008). Age-related differences in valence and arousal ratings of pictures from the International Affective Picture System (IAPS): Do ratings become more extreme with age? *Behavior Research Methods*, 40(2), 512–521.

Grundy, D., & Schemann, M. (2007). Enteric nervous system. *Current Opinion in Gastroenterology*, 23, 121–126.

Hajcak, G., Molnar, C., George, M. S., Bolger, K., Koola, J., & Nahas, Z. (2007). Emotion facilitates action: A transcranial magnetic stimulation study of motor cortex excitability during picture viewing. *Psychophysiology*, 44, 91–97.

Harmon-Jones, E., & Allen, J. J. B. (1998). Anger and frontal brain activity: EEG asymmetry consistent with approach motivation despite negative affective valence. *Journal of Personality and Social Psychology*, 74(5), 1310–1316.

Harmon-Jones, E., & Allen, J. J. B. (2001). The role of affect in the mere exposure effect: Evidence from psychophysiological and individual differences approaches. *Personality and Social Psychology Bulletin*, 27(7), 889–898.

- Harmon-Jones, E., & Peterson, C. K. (2009). Supine body position reduces neural response to anger evocation. *Psychological Science*, 20(10), 1209–1210.
- Harmon-Jones, E., & Sigelman, J. (2001). State anger and prefrontal brain activity: Evidence that insult-related relative left-prefrontal activation is associated with experienced anger and aggression. *Journal of Personality and Social Psychology*, 80(5), 797–803.
- Hennenlotter, A., Dresel, C., Castrop, F., Ceballos-Baumann, A. O., Wohlschläger, A. M., & Haslinger, B. (2009). The link between facial feedback and neural activity within central circuitries of emotion: New insights from botulinum toxin-induced denervation of frown muscles. *Cerebral Cortex*, 19(3), 537–542.
- Hitchcock, J. M., & Davis, M. (1987). Fear-potentiated startle using an auditory conditioned stimulus: Effect of lesions of the amygdala. *Physiology & Behavior*, 39, 403–408.
- Innes-Ker, A., & Niedenthal, P. M. (2002). Emotion concepts and emotional states in social judgment and categorization. *Journal of Personality and Social Psychology*, 83(4), 804–816.
- Ito, T. A., Cacioppo, J. T., & Lang, P. J. (1998). Eliciting affect using the International Affective Picture System: Trajectories through evaluative space. *Personality and Social Psychology Bulletin*, 24(8), 855–879.
- Izard, C. E. (1979). *The maximally discriminative facial movement coding system (MAX)*. Newark: University of Delaware Office of Instructional Technology.
- Izard, C. E., Dougherty, F. E., Bloxom, B. M., & Kotsch, N. E. (1974). *The Differential Emotions Scale: A method of measuring the meaning of subjective experience of discrete emotions*. Nashville, TN: Vanderbilt University.
- Jabbi, M., Swart, M., & Keysers, C. (2007). Empathy for positive and negative emotions in the gustatory cortex. *NeuroImage*, 34(4), 1744–1753.
- James, J., & Rogers, P. (2005). Effects of caffeine on performance and mood: Withdrawal reversal is the most plausible explanation. *Psychopharmacology*, 182(1), 1–8.
- James, W. (1890). *The principles of psychology*. New York: Holt.

- Jamieson, J. P., Koslov, K., Nock, M. K., & Mendes, W. B. (2013). Experiencing discrimination increases risk taking. *Psychological Science*, 24(2), 131–139.
- Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5), 770–814.
- Juslin, P. N., & Sloboda, J. A. (2001). *Music and emotion: Theory and research. Series in affective science*. New York: Oxford University Press.
- Kan, I. P., Barsalou, L. W., Solomon, K. O., Minor, J. K., & Thompson-Schill, S. L. (2003). Role of mental imagery in a property verification task: fMRI evidence for perceptual representations of conceptual knowledge. *Cognitive Neuropsychology*, 20(3), 525–540.
- Kassam, K.S., & Mendes, W.B. (2013). The effects of measuring emotion: Physiological reactions to emotional situations depend on whether someone is asking. *PLOS One*, 8(6), e64959.
- Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, 330(6006), 932.
- Kim, M. J., Loucks, R. A., Maital, N., Davis, F. C., Oler, J. A., Mazzulla, E. C., & Whalen, P. J. (2010). Behind the mask: The influence of mask-type on amygdala response to fearful faces. *SCAN*, 5, 363–368.
- Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The “Trier Social Stress Test” – a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1–2), 76–81.
- Kober, H., Barrett, L. F., Joseph, J., Bliss-Moreau, E., Lindquist, K., & Wager, T. D. (2008). Functional grouping and cortical-subcortical interactions in emotion: A meta-analysis of neuroimaging studies. *NeuroImage*, 42(2), 998–1031.
- Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, 84(3), 394–421.
- Kring, A. M., & Sloan, D. M. (1991). The Facial Expression Coding System (FACES): A users guide. Unpublished manuscript. Available at: <http://ist-socrates.berkeley.edu/%3C;akring/FACES%25;20manual.pdf>

- Kringelbach, M. L., & Rolls, E. T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology. *Progress in Neurobiology*, 72, 341–372.
- Kuppens, P., Tuerlinckx, F., Russell, J. A., & Barrett, L. F. (2012). The relation between valence and arousal in subjective experience. *Psychological Bulletin*, advance online publication. doi: 10.1037/a0030811.
- Lambie, J. A., & Marcel, A. J. (2002). Consciousness and the varieties of emotion experience: A theoretical framework. *Psychological Review*, 109(2), 219–259.
- Lambon Ralph, M. A., Pobric, G., & Jefferies, E. (2009). Conceptual knowledge is underpinned by the temporal pole bilaterally: Convergent evidence from rTMS. *Cerebral Cortex*, 19(4), 832–838.
- Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review*, 97(3), 377–395.
- Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical report A-8.
- Lange, K., Williams, L. M., Young, A. W., Bullmore, E. T., Brammer, M. J. *et al.* (2003). Task instructions modulate neural responses to fearful facial expressions. *Biological Psychiatry*, 53, 226–232.
- Leahy, R. M., Mosher, J. C., Spencer, M. E., Huang, M. X., & Lewine, J. D. (1998). A study of dipole localization accuracy for MEG and EEG using a human skull phantom. *Electroencephalography and Clinical Neurophysiology*, 107(2), 159–173.
- LeDoux, J. E., Iwata, J., Cicchetti, P., & Reis, D. J. (1988). Different projections of the central amygdaloid nucleus mediate autonomic and behavioral correlates of conditioned fear. *The Journal of Neuroscience*, 8(7), 2517–2529.
- Lench, H. C., Flores, S. A., & Bench, S. W. (2011). Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: A meta-analysis of experimental emotion elicitation. *Psychological Bulletin*, 137(5), 834–855.
- Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology*, 81(1), 146–159.

- Lewis, P. A., Critchley, H. D., Rotshtein, P., & Dolan, R. J. (2007). Neural correlates of processing valence and arousal in affective words. *Cerebral Cortex*, 17(3), 742–748.
- Libkuman, T. M., Otani, H., Kern, R. P., Viger, S. G., & Novak, N. (2007). Multidimensional normative ratings for the International Affective Picture System. *Behavior Research Methods, Instruments, & Computers*, 39, 326–334.
- Lieberman, M. D., Eisenberger, N. I., Crockett, M. J., Tom, S. M., Pfeifer, J. H., & Way, B. M. (2007). Putting feelings into words: Affect labeling disrupts amygdala activity in response to affective stimuli. *Psychological Science*, 18(5), 421–428.
- Lieberman, M. D., Inagaki, T. K., Tabibnia, G., & Crockett, M. J. (2011). Subjective responses to emotional stimuli during labeling, reappraisal, and distraction. *Emotion*, 11(3), 468–480.
- Lindquist, K. A., & Barrett, L. F. (2008). Constructing emotion: The experience of fear as a conceptual act. *Psychological Science*, 19(9), 898–903.
- Lindquist, K. A., & Gendron, M. (2013). What's in a word? Language constructs emotion perception. *Emotion Review*, 5(1), 66–71.
- Lindquist, K. A., Siegel, E. H., Quigley, K. S., & Barrett, L. F. (2013). The Hundred-Year Emotion War: Are emotions natural kinds or psychological constructions? Comment on Lench, Flores, and Bench (2011). *Psychological Bulletin*, 139(1), 255–263.
- Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences*, 35, 121–143.
- Lovullo, W. R., Wilson, M. F., Pincomb, G. A., Edwards, G. L., Topmkins, P., & Brackett, D. J. (1985). Activation patterns to aversive stimulation in man: Passive exposure versus effort to control. *Psychophysiology*, 22(3), 283–291.
- Lutz, C. (1985). Ethnopsychology compared to what? Explaining behaviour and consciousness among the Ifaluk. In G. M. White & J. Kirkpatrick (Eds.), *Person, self, and experience: Exploring Pacific ethnopsychologies* (pp. 35–79). Berkeley: University of California Press.
- Lynn, S. K., Zhang, X., & Barrett, L. F. (2012). Affective state influences

- perception by affecting decision parameters underlying bias and sensitivity. *Emotion*, 12(4), 726–736.
- Malle, B. F., & Holbrook, J. (2012). Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology*, 102(4), 661–684.
- Marks, D. F. (1973). Visual imagery differences in the recall of pictures. *British Journal of Psychology*, 64(1), 17–24.
- Matsumoto, D., & Ekman, P. (1988). *Japanese and Caucasian facial expressions of emotion and neutral faces (JACFEE and JACNeuF)*. Human Interaction Laboratory, University of California, San Francisco, 401.
- Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion*, 23(2), 209–237.
- McIntosh, D. N. (1996). Facial feedback hypotheses: Evidence, implications, and directions. *Motivation and Emotion*, 20, 121–147.
- Meeren, H. K. M., Van Heijnsbergen, C. C. R. J., & DeGelder, B. (2005). Rapid perceptual integration of facial expression and emotional body language. *Proceedings of the National Academy of Sciences*, 102(45), 16518–16523.
- Mehl, M. R., & Conner, T. S. (2012). *Handbook of research methods for studying daily life*. New York: Guilford Press.
- Michaliszyn, D., Marchand, A., Bouchard, S., Martel, M. O., & Poirier-Bisson, J. (2010). A randomized, controlled clinical trial of in virtuo and in vivo exposure for spider phobia. *CyberPsychology, Behavior, and Social Networking*, 13(6), 689–695.
- Mikels, J. A., Frederickson, B. L., Larkin, G. R., Lindberg, C. M., Maglio, S. J., & Reuter-Lorenz, P. A. (2005). Emotional category data on images from the International Affective Picture System. *Behavior Research Methods*, 37(4), 626–630.
- Morel, S., Ponz, A., Mercier, M., Vuilleumier, P., & George, N. (2009). EEG-MEG evidence for early differential repetition effects for fearful, happy and neutral faces. *Brain Research*, 13(1254), 84–98.
- Morrison, S. F. (2001). Differential control of sympathetic outflow. *American Journal of Physiology: Regulatory, Integrative and Comparative Physiology*,

281, R683–R698.

- Niedenthal, P. M. (2007). Embodying emotion. *Science*, 316(5827), 1002–1005.
- Niedenthal, P. M., Mermillod, M., Maringer, M., & Hess, U. (2010). The Simulation of Smiles (SIMS) Model: Embodied simulation and the meaning of facial expression. *Behavioral and Brain Sciences*, 33, 417–480.
- Norman, G. J., Cacioppo, J. T., Morris, J. S., Karelina, K., Malarkey, W. B., DeVries, A. C., & Berntson, G. G. (2011). Selective influences of oxytocin on the evaluative processing of social stimuli. *Journal of Psychopharmacology*, 25(10), 1313–1319.
- Norman, G. J., Hawkley, L. C., Cole, S. W., Berntson, G. G., & Cacioppo, J. T. (2011). Social neuroscience: The social brain, oxytocin, and health. *Social Neuroscience*, 1–12.
- Ochsner, K. N., Knierim, K., Ludlow, D. H., Hanelin, J., Ramachandran, T., Glover, G., & Mackey, S. C. (2004). Reflecting upon feelings: An fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience*, 16(10), 1746–1772.
- Olatunji, B. O., Babson, K. A., Smith, R. C., Feldner, M. T., & Connolly, K. M. (2009). Gender as a moderator of the relation between PTSD and disgust: A laboratory test employing individualized script-driven imagery. *Journal of Anxiety Disorders*, 23(8), 1091–1097.
- Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, 97, 315–331.
- Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The Measurement of Meaning*. Urbana: University of Illinois Press.
- Owren, M. J., Amoss, R. T., & Rendell, D. (2011). Two organizing principles of vocal production: Implications for nonhuman and human primates. *American Journal of Primatology*, 73, 530–544.
- Owren, M. J., & Bachorowski, J. (2007). Measuring emotion-related vocal acoustics. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment* (pp. 239–266). New York: Oxford University Press.
- Owren, M. J., & Rendell, D. (1997). An affect-conditioning model of nonhuman

- primate vocal signaling. In D. H. Owings, M. D. Beecher, & N. S. Thompson (Eds.), *Perspectives in ethology: Communication* (Vol. 12, pp. 299–346). New York: Plenum.
- Palermo, R., & Rhodes, G. (2007). Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia*, 45(1), 75–92.
- Panksepp, J. (1998). The periconscious substrates of consciousness: Affective states and the evolutionary origins of the self. *Journal of Consciousness Studies*, 5(5–6), 566–582.
- Patel, S., Scherer, K. R., Bjorkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. *Biological Psychology*, 87, 93–98.
- Paton, J., Boscan, P., Pickering, A., & Nalivaiko, E. (2005). The yin and yang of cardiac autonomic control: Vago-sympathetic interactions revisited. *Brain Research Reviews*, 49(3), 555–565.
- Peterson, C. K., & Harmon-Jones, E. (2012). Anger and testosterone: Evidence that situationally-induced anger relates to situationally-induced testosterone. *Emotion*, 12(5), 899–902.
- Philippot, P. (1993). Inducing and assessing differentiated emotion-feeling states in the laboratory. *Cognition & Emotion*, 7(2), 171–193.
- Pitcher, D., Garrido, L., Walsh, V., & Duchaine, B. C. (2008). Transcranial magnetic stimulation disrupts the perception and embodiment of facial expressions. *Journal of Neuroscience*, 28(36), 8929–8933.
- Pizzagalli, D. (2007). Electroencephalography and high-density electrophysiological source localization. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed., pp. 56–84). New York: Cambridge University Press.
- Powers, M. B., & Emmelkamp, P. (2008). Virtual reality exposure therapy for anxiety disorders: A meta-analysis. *Journal of Anxiety Disorders*, 22, 561–569.
- Price, T. F., & Harmon-Jones, E. (2010). The effect of embodied emotive states on cognitive categorization. *Emotion*, 10(6), 934–938.
- Prossin, A. R., Koch, A. E., Campbell, P. L., McInnis, M. G., Zalcman, S. S., &

- Zubieta, J.-K. (2011). Association of plasma interleukin-18 levels with emotion regulation and mu-opioid neurotransmitter function in major depression and healthy volunteers. *Biological Psychiatry*, 69(8), 808–812.
- Quigley, K. S., Barrett, L. F., & Weinstein, S. (2002). Cardiovascular patterns associated with threat and challenge appraisals: A within-subjects analysis. *Psychophysiology*, 39(3), 292–302.
- Quigley, K. S., & Berntson, G. G. (1990). Autonomic origins of cardiac responses to nonsignal stimuli in the rat. *Behavioral Neuroscience*, 104, 751–762.
- Riva, G., Mantovani, F., Capideville, C. S., Preziosa, A., Morganti, F. *et al.* (2007). Affective interactions using virtual reality: The link between presence and emotions. *CyberPsychology and Behavior*, 10(1), 45–56.
- Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, 128(6), 934–960.
- Ruby, P., & Decety, J. (2004). How would you feel versus how do you think she would feel? A neuroimaging study of perspective-taking with social emotions. *Journal of Cognitive Neuroscience*, 16(6), 988–999.
- Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, 115(1), 102–141.
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145–172.
- Russell, J. A., Bachorowski, J.-A., & Fernandez-Dols, J.-M. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology*, 54, 329–349.
- Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5), 805–819.
- Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, 57(3), 493–502.
- Schachter, J., & Singer, J. E. (1962). Cognitive, social, and physiological

- determinants of emotional state. *Psychological Review*, 69(5), 379–399.
- Schaefer, A., Nils, F., Sanchez, X., & Philippot, P. (2010). Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition & Emotion*, 24(7), 1153–1172.
- Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of Voice*, 9(3), 235–248.
- Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40, 227–256.
- Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 433–456). New York: Oxford University Press.
- Schmitt, B. M., Münte, T. F., & Kutas, M. (2000). Electrophysiological estimates of the time course of semantic and phonological encoding during implicit picture naming. *Psychophysiology*, 37(4), 473–484.
- Schutter, D. J. L. G., de Weijer, A. D., Meuwese, J. D. I., Morgan, B., & van Honk, J. (2008). Interrelations between motivational stance, cortical excitability, and the frontal electroencephalogram asymmetry of emotion: A transcranial magnetic stimulation study. *Human Brain Mapping*, 29, 574–580.
- Schwartz, G. E., Fair, P. L., Salt, P., Mandel, M. R., & Klerman, G. L. (1976). Facial muscle patterning to affective imagery in depressed and nondepressed subjects. *Science*, 192, 489–491.
- Simon-Thomas, E. R., Keltner, D., Sauter, D., Sinicropi-Yao, L., & Abramson, A. (2009). The voice conveys specific emotions: Evidence from vocal burst displays. *Emotion*, 9(6), 838–846.
- Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2008). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience*, 21(3), 489–510.
- Stepper, S., & Strack, F. (1993). Proprioceptive determinants of emotional and nonemotional feelings. *Journal of Personality and Social Psychology*, 64(2), 211–220.

- Stern, R. M., Ray, W. J., & Quigley, K. S. (2001). *Psychophysiological recording* (2nd ed.). New York: Oxford University Press.
- Stevenson, R. A., Mikels, J. A., & James, T. W. (2007). Characterization of the affective norms for english words by discrete emotional categories. *Behavior Research Methods*, 39(4), 1020–1024.
- Stevenson, R. J., Hodgson, D., Oaten, M. J., Barouei, J., & Case, T. I. (2011). The effect of disgust on oral immune function. *Psychophysiology*, 48(7), 900–907.
- Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54(5), 768–777.
- Tamir, M., & Ford, B. Q. (2009). Choosing to be afraid: Preferences for fear as a function of goal pursuit. *Emotion*, 9(4), 488–497.
- Tan, E. (2000). Emotion, art and the humanities. In J. M. Haviland-Jones & M. Lewis (Eds.), *Handbook of emotions* (2nd ed., pp. 116–136). New York: Guilford.
- Tassinary, L. G., & Cacioppo, J. T. (1992). Unobservable facial actions and emotion. *Psychological Science*, 3, 28–33.
- Tassinary, L. G., Cacioppo, J. T., & Geen, T. R. (1989). A psychometric study of surface electrode placements for facial electromyographic recording: I. The brow and cheek muscle regions. *Psychophysiology*, 26(1), 1–16.
- Tassinary, L. G., Cacioppo, J. T., & Vanman, E. J. (2007). The skeletomotor system: Surface electromyography. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed.). New York: Cambridge University Press.
- Teachman, B. A. (2007). Evaluating implicit spider fear associations using the Go/No-go Association Task. *Journal of Behavior Therapy and Experimental Psychiatry*, 38(2), 156–167.
- Tomaka, J., Blascovich, J., Kibler, J. L., & Ernst, J. M. (1997). Cognitive and physiological antecedents of threat and challenge appraisal. *Journal of Personality and Social Psychology*, 73(1), 63–72.
- Vazdarjanova, A., & McGaugh, J. L. (1998). Basolateral amygdala is not critical

- for cognitive memory of contextual fear conditioning. *Proceedings of the National Academy of Sciences USA*, 95(1), 5003–5007.
- Velten, E. (1968). A laboratory task for induction of mood states. *Behaviour Research and Therapy*, 6(4), 473–482.
- Vytal, K., & Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: A voxel-based meta-analysis. *Journal of Cognitive Neuroscience*, 22(12), 2864–2885.
- Wager, T. D., Barrett, L. F., Bliss-Moreau, E., Lindquist, K., Duncan, S. *et al.* (2008). The neuroimaging of emotion. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *The handbook of emotion* (3rd ed., pp. 249–271). New York: Guilford.
- Wager, T. D., Waugh, C. E., Lindquist, M., Noll, D. C., Fredrickson, B. L., & Taylor, S. F. (2009). Brain mediators of cardiovascular responses to social threat, Part I: Reciprocal dorsal and ventral subregions of the medial prefrontal cortex and heart-rate reactivity. *NeuroImage*, 47(3), 821–835.
- Watson, D., & Clark, L. A. (1994). *The PANAS-X: Manual for the Positive and Negative Affect Schedule – Expanded Form*. Iowa City: The University of Iowa Press.
- Weiskopf, N., Veit, R., Erb, M., Mathiak, K., Grodd, W., Goebel, R., & Birbaumer, N. (2003). Physiological self-regulation of regional brain activity using real-time functional magnetic resonance imaging (fMRI): Methodology and exemplary data. *NeuroImage*, 19, 577–586.
- Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B., & Jenike, M. A. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *The Journal of Neuroscience*, 18(1), 411–418.
- Williams, L. A., & DeSteno, D. (2008). Pride and perseverance: The motivational role of pride. *Journal of Personality and Social Psychology*, 94(6), 1007–1017.
- Wilson-Mendenhall, C. D., Barrett, L. F., Simmons, W. K., & Barsalou, L. W. (2011). Grounding emotion in situated conceptualization. *Neuropsychologia*, 49, 1105–1127.
- Winkielman, P., Berridge, K. C., & Wilbarger, J. L. (2005). Unconscious

affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Personality and Social Psychology Bulletin*, 31(1), 121–135.

Yeshurun, Y., & Sobel, N. (2010). An odor is not worth a thousand words: From multidimensional odors to unidimensional odor objects. *Annual Review of Psychology*, 61, 219–241.

Yik, M. S. M., Russell, J. A., & Barrett, L. F. (1999). Structure of self-reported current affect: Integration and beyond. *Journal of Personality and Social Psychology*, 77(3), 600–619.

Yoo, S.-S., & Jolesz, F. A. (2002). Junctional MRI for neurofeedback: Feasibility study on a hand motor task. *NeuroReport*, 13(11), 1377–1381.

Zuckerman, M., Klorman, R., Larrance, D. T., & Spiegel, N. H. (1981). Facial, autonomic, and subjective components of emotion: The facial feedback hypothesis versus the externalizer-internalizer distinction. *Journal of Personality and Social Psychology*, 41(5), 929–944.

Chapter eleven Complex Dynamical Systems in Social and Personality Psychology

Theory, Modeling, and Analysis

Michael J. Richardson, Rick Dale and Kerry L. Marsh

All social processes fundamentally involve change in time: Judgments materialize quickly over milliseconds or seconds, conversations flow over minutes, and relationships evolve across even longer time scales. Put simply, social systems are dynamical systems. The word “dynamical” simply means time-evolving and thus a *dynamical system* is simply a system whose behavior evolves or changes over time. Proposing that social processes are dynamical is not new and has a long history in social psychology (e.g., Asch, 1952; Lewin, 1936; Mead, 1934). Moreover, most researchers in social-personality psychology would agree that social processes and behavior are dynamical and change over time. Traditionally, however, social-personality psychology, like experimental psychology in general, has focused on summary statistics, which aggregate over time, such as in the form of magnitudes (e.g., behavioral frequencies, emotional intensities, and so on). This traditional approach is rooted in the linear statistical methods developed by Fisher and others (Meehl, 1978), and is aimed at detecting whether treatments or manipulations, on the whole, affect the outcome of some measured behavioral state or variable. Behavioral change is therefore conceptualized as the difference between static measures and is modeled by covarying responses on such measures. Unfortunately, this traditional approach merely describes behavioral change; it does not capture true time evolution and so is not always optimal for understanding the process by which behavioral change occurs. To make progress in our understanding of psychological *change* and *process*, therefore, researchers need to consider adopting new tools and methodological concepts, namely those of dynamical systems.

The scientific study of dynamical systems is concerned with understanding, modeling, and predicting the ways in which the behavior of a system changes over time. As a formal approach, it has a long history in applied mathematics and physics, and has been used extensively to understand and model the

behavior of many different types of physical systems, such as the motion (position and velocity) of planets, mass-spring systems, swinging pendulums, and self-sustained oscillators. In the last few decades, however, an increasing number of researchers have begun to investigate and understand the dynamic behavior of more *complex* biological, cognitive, and social systems, using the concepts and tools of dynamical systems.

The term “complexity” refers to the fact that most biological, cognitive, and social systems typically exhibit behavior that is nonlinear and involves a large number of interacting elements or components. Historically, it is the nonlinearity of complex dynamical systems that has largely hindered research on such systems, in that the numerical techniques that enable one to uncover the dynamics of nonlinear and complex dynamical systems involve an extensive number of computational processes that are impossible to perform without modern computers. This is true for both abstract nonlinear dynamical models (covered in the second section of this chapter) and for the analysis of behavioral data (discussed in the third section). These days, of course, these difficulties of computation no longer exist, and researchers can formulate and analyze many nonlinear and complex dynamical systems quite easily. Indeed, the fields of nonlinear dynamics and complex systems, as well as our theoretical understanding of such systems, have grown in parallel with increases in the availability and computational power of modern computers.

Advances in the modeling and analysis of complex dynamical systems have also led to the steady rise of dynamical systems in social-personality psychology. Vallacher, Read, and Nowak (2002) offer an excellent review of this rise and describe how dynamical theory and models can be employed to characterize an array of socially relevant behaviors, such as attitude change, social judgment, and self-perception (for reviews, see Nowak & Vallacher, 1998; Vallacher & Nowak, 1994). This interest in dynamics and the dynamical systems approach has continued to expand in social-personality psychology, with numerous researchers embracing a dynamical understanding of mood (e.g., Gottschalk, Bauer, & Whybrow, 1995; Schuldborg & Gottlieb, 2002), personality expression (Brown & Moskowitz, 1998; Mischel & Shoda, 1995), conformity (Tesser & Achee, 1994), romantic relationships (Gottman, Swanson, & Swanson, 2002), and person perception and construal (Freeman & Ambady, 2011). Along with more recent advances in fractal methods (Correll, 2008; Delignières, Fortes, & Ninot, 2004), decision dynamics (Freeman, Dale, & Farmer, 2011), cognitive dynamics (Spivey, 2007), coordination dynamics (Kelso, 1995; Schmidt & Richardson, 2008), and the behavioral dynamics of perception and action (

Warren, 2006), the concepts and tools that have been developed for understanding complex dynamical systems appear to provide a promising new method for understanding social behavior and cognition.

Here we provide an introductory overview of the dynamical systems approach and how the theoretical concepts, modeling techniques, and analysis tools used to investigate complex dynamical systems can be used to understand social behaviors that emerge and change over time. It is by no means a comprehensive review of complex dynamical systems and the dynamical systems approach. Rather, the chapter is aimed at providing the reader with the knowledge needed to seek out more detailed discussions elsewhere. Throughout the chapter we include key references that provide a deeper and more detailed understanding of the concepts and issues discussed.

The chapter is divided into three main sections. The first section covers the basic concepts of dynamical systems theory, including their implications for understanding human and social behavior, and how complex ordered behavior emerges from the nonlinear interactions that exist between the components of a behavioral system. The second section covers the basic forms and mathematical properties of dynamical systems models and how dynamical modeling can provide insights about the organization or reorganization of stable (or unstable) human and social behavior. The final section addresses the basic aspects of dynamical systems analysis and focuses on a number of analysis techniques that can be employed to uncover the dynamics of a behavioral system from time-series recordings.

Dynamical Systems Theory

The keystone concept throughout this section (and the entire chapter) is that the behavior of a complex dynamical system can arise in a self-organized manner from the free interplay of components and properties of the system. To unpack what this means, we begin by briefly defining what a complex dynamical system is and then go on to describe the abstract properties that underlie the behaviors that complex dynamical systems exhibit.

What Is a Complex Dynamical System?

The term “complex dynamical system” lacks a consensus definition, but many researchers agree that complex dynamical systems exhibit three key characteristics (see Gallagher & Appenzeller, 1999, and articles therein, for a

discussion). First, they consist of a large number of interacting components or agents. This may be said about the behavior of an individual person, as in the interacting constraints that drive social perceptions and decisions (e.g., Read & Miller, 2002; Smith, 1996). It may also be said about the behavior of groups of human beings, such as in dyadic conversation (Buder, 1991) or small work-teams (Arrow, 1997). A second property is that these systems exhibit emergence: Their collective behavior can be difficult to anticipate from knowledge of the individual components that make up the system, and exhibit some coherent pattern or even, in some cases, apparent purposiveness. For example, a person's social judgment may be a nonobvious consequence of the interaction among an array of informational constraints (e.g., Freeman & Ambady, 2011), and group behaviors may be a nonobvious outcome of interactions among group members (Arrow, 1997). Third, this emergent behavior is self-organized and does not result from a central or external controlling component process or agent. Although all three characteristics are necessary for a system to be considered complex, the appearance of emergent behavior that results from self-organization is the most distinguishing feature of complex systems (Boccaro, 2003). Accordingly, we turn next to the topic of self-organization.

Self-Organization

The term “self-organization” is used to refer to behavioral patterns that emerge from the interactions that bind the components of a system (social or otherwise) into a collective, synergistic system, while not being dictated a priori by a centralized controller. As mentioned earlier, self-organization is synonymous with complex systems, and interestingly enough, the most widely used examples of a self-organized and emergent behavioral process are social examples, in particular the coordinated activities of social insects such as ants. Take a colony of harvester ants, for example, in which different members of the ant colony perform one of several different tasks: foraging, patrolling, nest maintenance, or midden work (clearing up debris). Individual ants do not perform the same task all the time, but transition between the different modes of behavior as the need arises, with the appropriate number of ants (workers) engaged in a particular task at any given time. That is, patrollers become foragers, foragers become nest maintenance workers, and midden workers becoming patrollers, and so forth, based on current conditions (e.g., Gordon, 2007). Task allocation, however, is not achieved by a centralized controller or leader. The queen does not decide who does what, nor does any other ant. In fact, it would be impossible for any

ant to oversee the entire colony. Rather, individual ants can only detect local tactile and chemical information, with the coordinated behavior of the colony emerging from the local interactions that constrain the tasks an individual ant should perform (Boccaro, 2003).

The harvester ant colony highlights how coordinated social behavior can result spontaneously from the physical and informational interactions of perceiving-acting agents. The nest-building behavior of other social insects, such as termites, bees, and wasps, is similarly self-organized. So too is the coordinated behavior of schools of fish, flocks of birds, and herds of ungulate mammals: No individual animal has precise control over the direction or behavior of the group (e.g., Camazine, Deneubourg, Franks, Sneyd, Theraulaz, & Bonabeau, 2001; Theraulaz & Bonabeau, 1995).

Though it is certainly different from that of humans in numerable respects, the self-organization that ants and other social animals promote is directly relevant to, and sometimes motivates, the study of human social groups (Arrow, 1997). For instance, the coordinated activity observed between pedestrians walking down a crowded sidewalk is the result of self-organizing dynamics (Sumpter, 2010). Goldstone and colleagues have demonstrated that the human-path systems that are created between regularly visited destinations (say, buildings around a university campus) emerge in a self-organized manner and are often mutually advantageous to the members of the group that created them. Task subroles and divisions of labor can also emerge and be spontaneously adopted by individuals during joint action or a social cognitive task (Eguíluz, Zimmermann, Cela-Conde, & San Miguel, 2005; Goldstone & Gureckis, 2009; Richardson, Marsh, & Baron, 2007; Theiner, Allen, & Goldstone, 2010). Even group dynamics and performance can be self-organized by means of integrative complexity, whereby the individual ideas and opinions of group members emerge and become dynamically integrated over time (Cummings, Schlosser, & Arrow, 1996).

Soft-Assembly

The kinds of self-organized dynamical systems described in the preceding section, such as social insects and sometimes groups of people, are temporary organizational structures that are put together in a fluid and flexible manner. In the case of the harvester ants, it does not matter which particular ant does which particular job; each ant is capable of taking up any job at any point in time. Though obviously different in important ways, humans also engage in modes of

behavior flexibly throughout the course of a single day (see Iberall & McCulloch, 1969 for classic a discussion; see Isenhower, Richardson, Marsh, Carello, & Baron, 2010 for a task example). Moreover, although it seems individuals in a group or social network are in complete control of their own behavior and are consciously aware of their acts (and could verbalize them if asked), we know that such centralized, conscious control is often an illusion: Classic research in social psychology suggests that individuals can be unaware of the “true” reasons for their actions (Nisbett & Wilson, 1977). Indeed, the coordinated behavior and intentions of socially situated individuals can be constrained and self-organized by environmental and situational constraints of which we are not aware, with the unfolding dynamics of human behavior reflecting a mutuality of responsiveness between individuals and the context in which they are embedded (Reis, 2008; Richardson, Marsh, & Schmidt, 2010; Semin & Smith 2008; Thelen & Smith, 1994). Understanding and predicting the time-evolving behavior of an individual or social system is therefore not only dependent on identifying the environmental factors or agents that make up the behavioral system in question, but also on the relevant interagent and agent-environment interactions that shape behavior. This implies that the dynamical behavior of cognitive and social systems is highly context dependent.

Dynamical systems that exhibit this kind of emergent, context-dependent behavior are often referred to as *softly assembled* or *soft-molded* systems, in that the behavioral system reflects a temporary coalition of coordinated entities, components, or factors. The term “synergy” is sometimes used to refer to softly assembled systems – a functional grouping of structural elements (molecules, genes, neurons, muscles, limbs, individuals, etc.) that are temporarily constrained to act as a single coherent unit (Kelso, 2009). In contrast, most nonbiological systems or machines are hard-molded systems. A pendulum clock, for example, is a hard-molded dynamical system, in that it is composed of a series of components (parts), each of which plays a specific, predetermined, and unchanging role in shaping the motion of the clock's pendulum over time. Rigidly organized factory assembly lines, organizations, or military structures could also be considered intuitive examples of hard-molded machine-like social systems. But even within rigid social structures one finds differing degrees of fluidity and soft-assembly, both within and across different levels of organization. One could argue that this soft-assembly, present to some degree, is a necessary requirement for the ongoing existence and success of such social systems and organizations (e.g., Dooley, 1994; Guastello, 2002).

Interaction-Dominant Dynamics

The key property of softly assembled systems is that they exhibit interaction-dominant dynamics, as opposed to component-dominant dynamics. For component-dominant dynamical systems, system behavior is the product of a rigidly delineated architecture of system modules, component elements, or agents, each with predetermined functions (i.e., the pendulum clock or a factory assembly line). As noted earlier, however, for softly assembled interaction-dominant dynamical systems, system behavior is the result of *interactions between* system components, agents, and situational factors, with these intercomponent and interagent interactions altering the dynamics of the component elements, situational factors, and agents themselves (Anderson, Richardson, & Chemero, 2012; Kello, Beltz, Holden, & Van Orden, 2007; Van Orden, Kloos, & Wallot, 2011).

Thus, if one were to examine the relationship between any two levels of an interaction-dominant dynamical system, one would observe that elements or agents at the lower level of the system modulate the macroscopic order of the higher level and at the same time are structured by the macroscopic order of the system. For example, the individuals (micro-level) within a cultural system (macro-level) modulate the behavioral order of the cultural system, with the dynamical organization of the cultural system in turn (and at the same time) modulating and structuring the behavior of the individuals within it. Accordingly, for interaction-dominant systems, it is difficult – and often impossible – to assign precise causal roles to particular components, factors, agents, or system levels. Of particular significance for the study of cognitive and social behavior is the implication that one cannot hope to appropriately understand behavioral organization by attempting to study systems components or agents in isolation. For that reason, researchers who have adopted the complex dynamical approach to social phenomena have started to conceptualize many domains of human behavior, from language acquisition (Van Geert, 1991) to group dynamics (Arrow, McGrath, & Berdahl, 2000), as being guided and self-organized by the dynamic interaction of many constraints, factors, and processes.

Nonlinearity

A nonlinear system is one in which the system's output is not directly proportional to the input, as opposed to a linear system in which the output can be simply represented as a weighted sum of input components. Complex

dynamical systems, most notably biological and social systems, are nonlinear in this sense. Our attraction to another individual or our self-concept, for instance, may not correspond to a mere average or sum of positive and negative attributes (Rinaldi & Gragnani, 1998; Sprott, 2004), but rather some form of multiplicative combination of attributes and situational factors that results in an attraction or self-concept that is more (or less) than the sum of its parts. The consequence of this principle for understanding human behavior and social systems is equivalent to the consequence of a system exhibiting interaction-dominant dynamics – a system's behavior cannot be reduced to a set of component dominant factors that interact in a simple linear fashion (Van Orden, Holden & Turvey, 2003). Thus, in the context of complex dynamical systems, the term “nonlinear” also means something more than multiplicative. More generally, it is used to refer to the non-decomposability of a complex dynamic system, whereby a nonlinear dynamical system is a system whose macroscopic behavioral order results from the complex interactions of micro-scale components, but via processes of circular causality, also modifies the interactions between micro-scale components as well as the behavior of the components themselves.

On the one hand, the nonlinearity of complex dynamical systems makes them much more difficult to understand. In fact, the effects of nonlinear processes are typically not known or cannot be known ahead of time. On the other hand, it is only because complex dynamical systems are nonlinear that they can exhibit complex emergent behavior. It is for these reasons that an increasing number of researchers and theorists consider human behavior and social processes to be predominantly nonlinear (see e.g., Guastello, Koopmans, & Pincus, 2009; Kelso, 1995; Vallacher & Nowak, 1994; Nowak & Vallacher, 1998 for edited collections and reviews). In fact, a defining feature of human behavior is that it is often unpredictable. It is this aspect of human behavior that makes the science of social-personality psychology, as well as the many other fields of psychology (perceptual, cognitive, clinical, etc.), so difficult and at the same time so intriguing. In addition, although the word “random” is often used in everyday speech when referring to the unpredictability of human behavior, the general belief that human behavior is not random, but rather complex, is also consistent with the notion of nonlinearity.

Chaos

The fact that nonlinear dynamical systems can exhibit complex and

unpredictable behavior is interesting in and of itself and highly relevant for the study of human behavior. Even more interesting, however, is one of the key discoveries from the study of nonlinear dynamical systems: that highly complex behavior can even emerge from very simple rules or systems so long as the components or agents of the system interact in a nonlinear manner. That is, very simple deterministic nonlinear systems can produce extremely complex and unpredictable behavior. One notable form of complex behavior they produce is *chaotic behavior*.

A classic example of a simple nonlinear system that results in chaotic behavior is the *logistic map*. The logistic map is a discrete dynamical system, meaning its behavior changes over discrete rather than continuous time steps (see next section for more details). More precisely, it is simple dynamical equation of the form

$$x_{(t+1)} = rx_{(t)}(1 - x_{(t)}) \quad (11.1)$$

where x is the behavioral variable, r is a fixed behavioral parameter, and t equals time from step 0 to step n (i.e., $t = 0, 1, 2, \dots, n$). To help make this equation less abstract and easier to understand, let us assume that this equation is used to model the daily mood of an individual diagnosed with bipolar disorder or manic depression, where x represents the individual's daily mood on a scale of 0 to 1, with 1 corresponding to a perfectly positive mood, and r representing the severity of the individual's diagnosis on a scale of 0 to 4, with 4 corresponding to a severe diagnosis. The predicted daily mood of the individual (i.e., $x_{(t+1)}$) is therefore a simple mathematical function of the current day's mood, $x_{(t)}$, multiplied by the severity of the diagnosis, r , multiplied by 1 minus the current day's mood, or $1 - x_{(t)}$. For illustrative purposes, imagine that an individual's diagnosis was relatively severe, that r equaled 3.8, and the person's initial mood of x on day zero equaled 0.6. If we then computed or iterated the equation 100 times, we would get the time-evolving behavioral pattern displayed in [Figure 11.1](#), where the value of x , or mood in our example, is plotted as a function of time step from 1 to 100. The complex and unpredictable nature of the behavioral pattern over time is quite evident, with the value of x from one time step to the next seeming to change in a way that far exceeds the simplicity of the equation that was used to generate it.

It is the latter feature of chaotic systems like the logistic map (Equation 11.1) that often surprises researchers. That is, while the behavior of such systems is

completely determined by a set of simple deterministic (i.e., nonrandom) equations or rules, it can be very complex and very difficult to predict. This is because chaotic systems exhibit a *sensitive dependence on initial conditions*: minor differences in starting states can become amplified as the system evolves over time. Given that measuring or knowing the initial conditions of any natural dynamical system with perfect precision is impossible, it is therefore equally impossible to predict the long-term behavior of such systems if they are chaotic.

The implication of this principle for understanding human social behavior is quite profound. Most obvious is that the existence of chaos forces researchers to truly reconsider what it means for a behavioral event to be random (Guastello & Liebovitch, 2009). Perhaps less obvious is that the apparent randomness of variable or noisy behavior (i.e., normally distributed random noise or variance) that is traditionally assumed in nearly all psychology studies becomes an empirical question. Just because a variable behavior appears to be random does not mean that one should conclude that it is random. Chaos also highlights the nonobvious connection between past and future events, and that even extremely trivial changes can have a significant effect on the outcomes of time-evolving behavior. It is for these reasons that chaos has found its way into extensive theoretical discussion of social psychological systems. A lucid introduction and review of this point is found in Barton (1994). One prominent application described in Barton's review is in the clinical realm, where chaos has enjoyed a rapid growth of application across a range of traditions. Another popular, visual introduction to chaos and nonlinear dynamics that utilizes clinical psychology is provided by Abraham, Abraham, and Shaw (1990).

Dynamical Systems Modeling

The goal of many research endeavors is to effectively model a behavioral system in order to make specific predictions about how the system will behave in the future. In personality and social psychology, this has typically been done using various forms of regression analysis and structural equation modeling (for instance, see Fabrigar and Wegener, Chapter 19 in this volume). An advantage of dynamical models is that they can handle the time-dependence of behavior and are not restricted to making linear assumptions about behavioral organization. Accordingly, nonlinear dynamical models can provide deep insights about the behavior of real-world time-evolving processes and can play a significant role in theorizing about how and why certain behavioral processes might emerge. For example, Meadows (2008) offers a lucid discussion of the

benefits of dynamical systems modeling for understanding the interactions and behavioral nonlinearities that anchor social systems, from the parameters that lead to stable sustenance of romantic relationships (Gottman, Murray, Swanson, Tyson, & Swanson, 2003) to the way that societal rules can change patterns of self-organization.

It is important to keep in mind that any model of a biological, psychological or social system is at best an idealization (this is just as true for dynamical models as for non-dynamical models) and that the goal of a dynamical model is to capture the most important features of a system or process (Bertuglia & Vaio, 2005). Dynamical modeling involves describing how the behavioral state of a system changes over time. Here the term “state” refers to the current value of a variable (or variables) that are used to capture the system in question. A state variable could be any property of a system that might change over *time*, such as the movements of the body, limb, or eyes during social interaction, or the mood, attitudes, or personality characteristics of a child or adult. One could even model the change in state of two people, such as the quality of a romantic relationship. Essentially, a dynamical model describes how the state variables of a system evolve over time through rules or equations that determine the system's future state given its current state.

Difference Equations

A difference equation is a *recursive function* sometimes called an *iterative map* and can be used to model the behavior of a system at discrete time steps (1, 2, 3... t , where t equals the number of time steps), with the state of the system at each time step defined as a function of the preceding state. The logistic map, introduced in Equation 11.1 in the previous section, is an example of a difference equation. More precisely, the logistic map is a one-dimensional nonlinear¹ difference equation – the dimension of a dynamical system equals the number of state variables needed to completely describe the system (i.e., one-, two-, three-,... to n -dimensions) – and is one of the most well-known and studied difference equations in the field of nonlinear dynamics. This is because despite the simplicity of Equation 11.1, it exhibits a wide variety of dynamic behaviors for different values of the system parameter r (specifically for $0 < r < 4$). This includes various types of stable fixed point and periodic behaviors, and, as we have already seen, even chaotic behavior (see Figure 11.2).

Of more relevance to understanding social behavior, Nowak and Vallacher have demonstrated how two coupled logistic equations can be used to model the

behavioral synchronization of two individuals in social interaction (e.g., Nowak & Vallacher, 1998; Nowak, Vallacher, & Borkowski, 2002; Vallacher, Nowak & Zochowski, 2005; also see Buder, 1991). Their model takes the form,

$$\begin{aligned} x_{1(t+1)} &= \frac{r_1 x_{1(t)} (1 - x_{1(t)}) + \alpha r_2 x_{2(t)} (1 - x_{2(t)})}{1 + \alpha} \\ x_{2(t+1)} &= \frac{r_2 x_{2(t)} (1 - x_{2(t)}) + \alpha r_1 x_{1(t)} (1 - x_{1(t)})}{1 + \alpha} \end{aligned} \quad (11.2)$$

where x is a generic variable representing the intensity of some observable, communicative behavior, such as approach (or avoidance), and r corresponds to internal states, such as personality traits, moods, attitudes, and values, that shape an individual's behavior over time. This is a two-dimensional model, in that there are two state variables, x_1 and x_2 , one state variable and equation for each individual's behavior. The equations are also *coupled*, in that the behavioral state of each individual is dependent on his or her own previous state, as well as the previous state of his or her partner, with the parameter α (alpha) determining the strength of the coupling (i.e., mutual influence). Although a detailed discussion of this model and the behaviors it generates is beyond the scope of the chapter, its significance is that it predicts increased social monitoring and mutual influence (i.e., communication, mutual reinforcement and self-and other-monitoring) when individuals with different internal states (e.g., personality traits, moods, attitudes) are required to synchronize their behavior. This increase in social monitoring and mutual influence is then presumed to put greater stress on the interactional system by reducing the executive resources. In contrast, when partners have similar internal states, behavioral synchronization can occur with little social monitoring or mutual influence, leaving more energy and cognitive resources available for the coupled individuals to pursue common goals.

Differential Equations

In contrast to difference equations, which model the behavior of state variables across discrete time steps, a *differential equation* is a mathematical equation that models the *continuous time evolution* of a system in terms of the *rate of change* of state variables over time. As a starting example, consider the simple one-dimensional nonlinear differential equation

$$\dot{x} = rx(1 - x) \quad (11.3)$$

where x is the state variable, \dot{x} is the rate of change of the x over time, and r is a state parameter. This equation is very similar to Equation 11.2 and is the logistic equation in differential form. However, because Equation 11.3 models the rate and direction in which x changes over continuous time, the way the value of x changes is quite different from that determined by Equation 11.2. Imagine that x represents some observable behavior, such as *attraction* – in the context of social interaction, attraction could refer to the pull toward another person, affiliation, or liking – and r corresponds to the number of positive attributes. If we restrict our consideration to initial conditions, $x_{(0)} > 0$, and $r > 0$, then x always approaches $x = 1$, no matter what initial condition we choose. In other words, setting the parameter $r > 0$ would always predict the same eventual level of attraction. This can be seen from an inspection of [Figure 11.3](#), in which the change in x over time is presented for four different initial conditions ($x_{(0)} = .1, .8, 1.6, 3.9$). Notice also that increasing r only changes the rate at which x approaches 1. It does not change the dynamics qualitatively, as it did for the logistic map in Equation 11.1. Because x is always attracted toward 1, irrespective of initial condition, $x = 1$ is the stable fixed point for Equation 11.3. We describe stable *fixed point attractors*, as well as other types of attractors, in more detail later in the chapter. At this stage it is sufficient to say that a stable fixed point is a state toward which the variables of a dynamical system move over time.

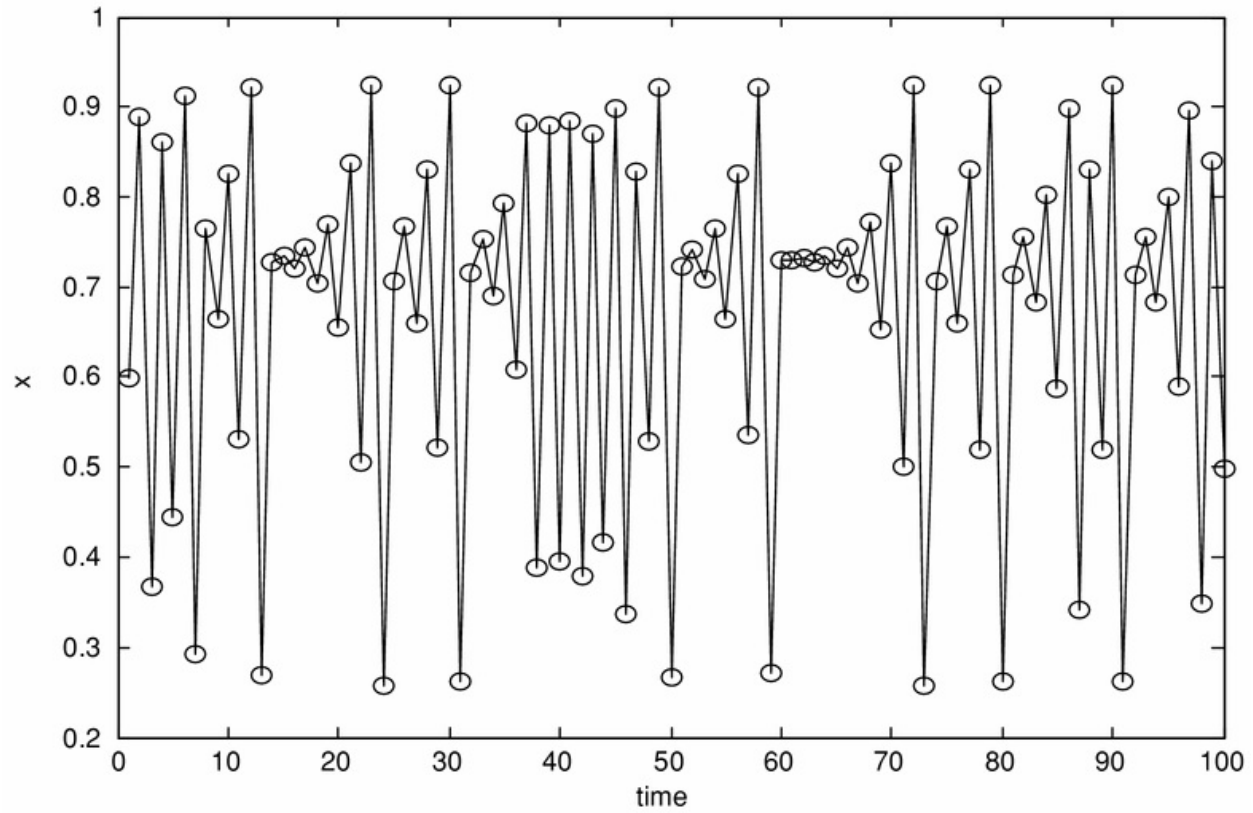


Figure 11.1. Behavior of the logistic map's dependent variable, x , during a chaotic regime (i.e., $r = 3.8$).

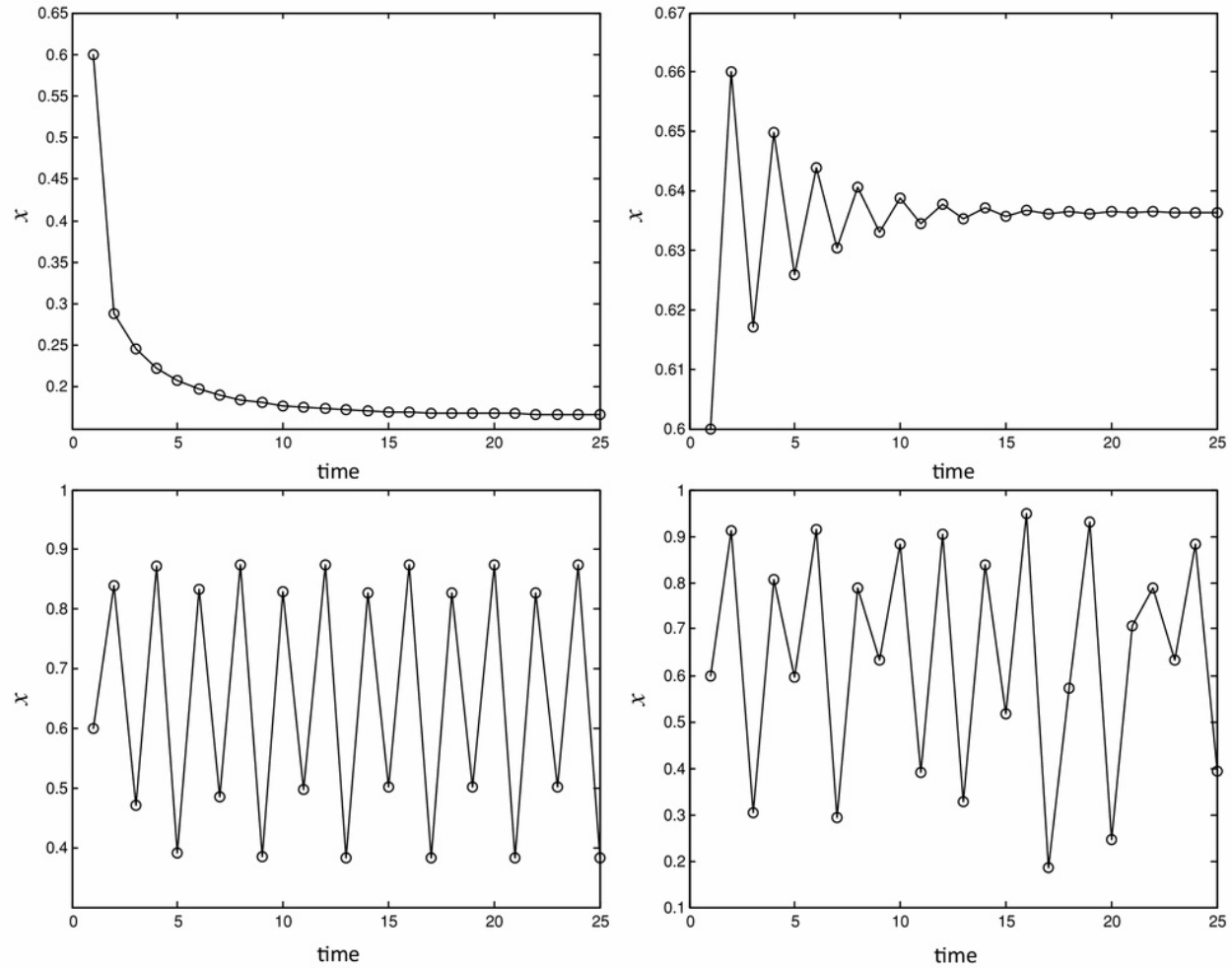


Figure 11.2. Examples of how x changes over time for the logistic map (Equation 11.1) for different values of r , but the same initial condition of $x_{(0)} = .6$. (top left) monotonic approach towards a fixed value of x for $r = 1.2$. (top right) oscillatory approach towards a fixed value of x for $r = 2.7$. (bottom left) periodical oscillates between 4 values of x for $r = 3.2$. (bottom right) chaotic behavior for $r = 3.8$.

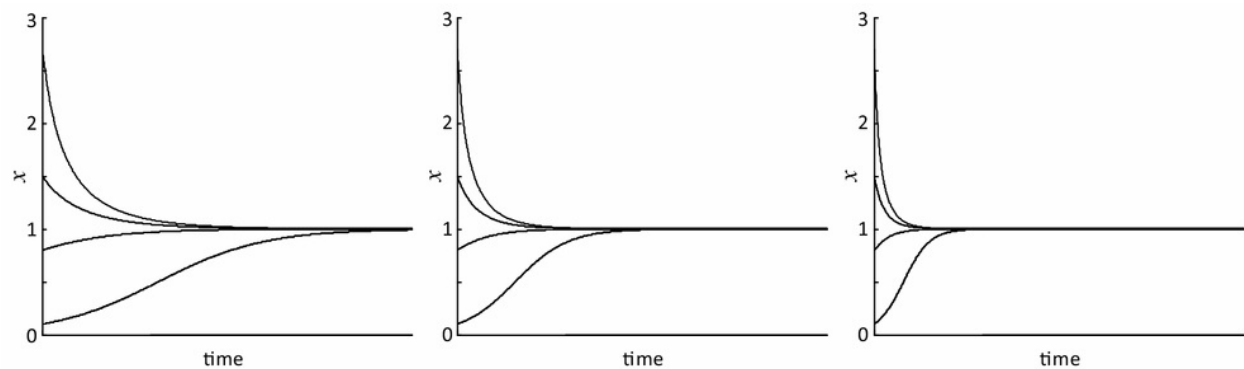


Figure 11.3. Examples of how x changes over time for Equation 11.3, using

four different initial conditions ($x_{(0)} = .1, .8, 1.6, 3.9$). For the left, middle, and right graphs, $r = .5, 1$, and 2 , respectively.

Although the simplicity of Equation 11.3 provides a good introduction to differential equations, the fact that its state variable, x , is always attracted toward 1 means that the degree to which this equation could be used to model complex human and social behavior is extremely limited. An example of a differential equation that is more relevant to understanding human and social behavior is

$$\dot{x} = k + x - x^3 \quad (11.4)$$

where x might represent an individual's attitude toward a certain political or racial group X and k is a state parameter that captures the amount of positive or negative experiences or information an individual has about group X . What is interesting about this system is that the number and location of its stable fixed points change as we change the parameter k . This is best illustrated by plotting Equation 11.4 as a potential function where the fixed points of the system are represented as wells in a one-dimensional landscape, with the depth of the well corresponding to the strength or stability of the stable fixed points or system attractors (see [Figure 11.4](#)).

Of particular relevance is that Equation 11.4 predicts that an individual's attitude toward group X would remain relatively stable, either negative or positive, across a range of k values and then at certain values of k suddenly transition from a negative to a positive state or from a positive to a negative state. More specifically, if an individual starts out with a negative attitude toward group X , and then k is increased from -1 to 1 (i.e., an individual starts to have more and more positive experiences with group X or receive more and more positive information about group X), the individual's attitude will be predicted to remain negative even beyond the point (i.e., $k > 0$) at which the individual has more positive than negative experiences with group X . The individual will finally transition to having a positive attitude toward group X after a critical number of positive experiences have occurred, that is, at $k = .35$. Conversely, if k is decreased from 1 to -1 , a transition from a positive to a negative attitude will occur at $k = -.35$.

The kind of sudden transition predicted by Equation 11.4 is called a phase transition, where *phase* refers to a qualitative state of the system (i.e., a negative

state or phase vs. a positive state or phase). Such transitions are a prominent characteristic of complex dynamical systems, and the fact that nonlinear differential equation models can be used to capture such behavior is part of their appeal. As the attitude example suggests, systems like Equation 11.4 are well suited for modeling qualitative changes in human cognition and social behavior. For instance, similar nonlinear systems can be employed to model transitions in categorical speech perception (Tuller, 2004; Tuller et al., 1994) and between conciliation and aggression during conflict situations (Colman, Vallacher, Nowak, & Bui-Wrzosinska, 2007).

Equations 11.2 and 11.3 are examples of one-dimensional differential equations because they have one state variable. Like difference equations, differential systems might also require more than one state variable to describe the behavior of the system. For instance, the motion of a pendulum requires a two-dimensional or second-order model that determines both its position and velocity over time. Differential systems with two or more state variables can also be coupled. For example, a two-dimensional system of coupled differential equations could be used to represent the population of two codependent species of animal (i.e., rabbits and foxes). With respect to personality and social psychology, two-dimensional systems have been used to model the coordinated behavior of interacting individuals (e.g., Baron et al., 1994; Schmidt & Richardson, 2008; Tesser & Achée, 1994) and the long-term marital success of a husband and wife (Gottman et al., 2002; Gottman et al., 2003).

Attractors

One of the key aims of dynamical modeling is to effectively capture the attractors of a system. An attractor is a state or subset of states toward which the dynamical system moves over time (i.e., corresponds to a final future state or set of states). Attractors are often described geometrically, such as a point or a closed curve, and can be intuitively visualized by plotting the time-evolving behavior of a dynamical system in its phase space. A *phase space* is the set of all possible states of a dynamical system, with each state corresponding to a unique point in phase space. A graphical depiction of a system's attractors and the set of possible trajectories that can be exhibited by a system given its attractor layout is called a *phase portrait* (see Figure 11.5).

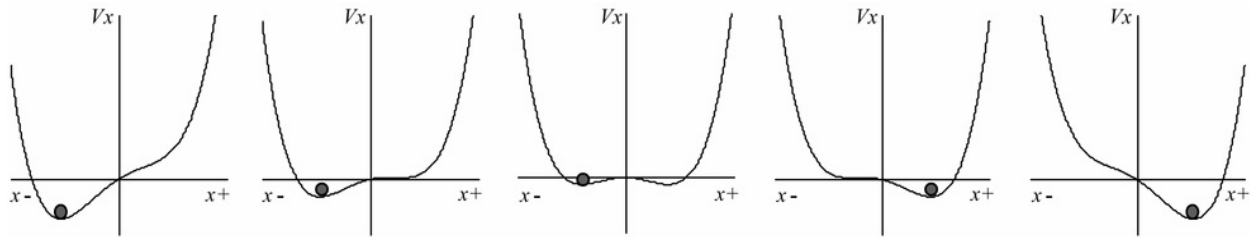


Figure 11.4. Potential functions for Equation 11.4, demonstrating how the system's stable fixed points change as k is scaled from -1 to 1 (left to right, respectively). In these plots, the x -axis corresponds to the possible values of the state variable, x , and the y -axis corresponds to the potential (Vx) of the system state to move to another state. Wells in the potential function or local minima (minimal potentials) correspond to stable fixed points, such that the state of the system (i.e., illustrated as a small grey ball) is trapped at the bottom of the well until that state is no longer stable (no longer a minima).

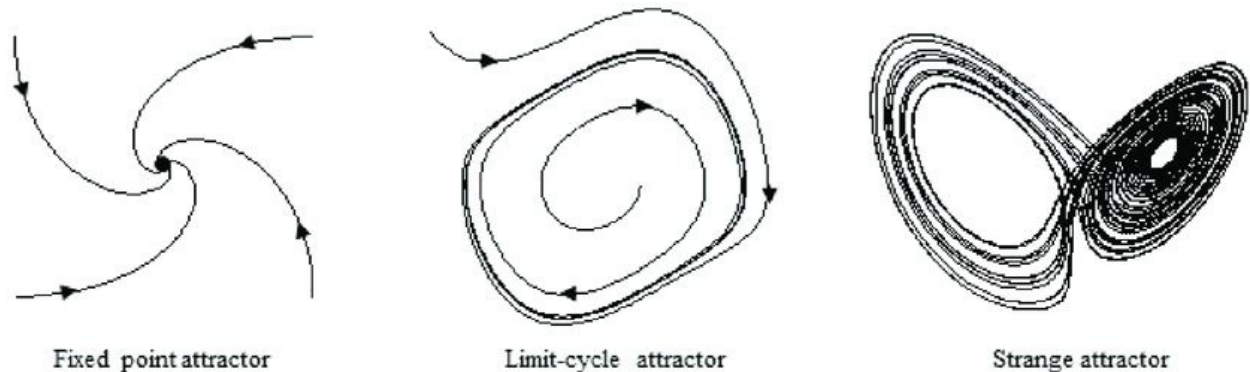


Figure 11.5. Prototypical examples of (left) a stable fixed point attractor, (middle) a limit cycle attractor, and (right) a strange attractor (see text for details).

The attractor concept has been fruitfully applied in many social contexts. For example, Gottman *et al.* (2002) have developed models in which the stabilities of intimate relationships can be captured using fixed point attractors. As defined earlier, a *fixed point attractor*, commonly referred to as a *stable fixed point* (sometimes called a stable node, equilibrium point, or point attractor), is a location in phase space that system trajectories converge toward as time increases (Figure 11.5 left). A system may have two or more stable fixed points, in which case the system is considered to be *multistable*. The models developed by Gottman and colleagues often have two fixed point attractors: sustained marriage or divorce. The authors design differential equations much like Equation 11.4 and explore the parameters that govern the dynamic regimes of a

married couple into one of these two equilibrium states. Although linear models can be applied in marriage research, the goal of the Gottman and colleagues' research agenda is to articulate the dynamic change – the trajectory that a couple may take – into one or another stable attractor (divorce, marital misery, happiness, etc.).

Space limits our discussion of all of the different types of attractors, but it is important to note two other types of attractor often featured in dynamical systems research, namely limit cycle and strange attractors. A *limit cycle attractor* is a subset of states within a system's phase space that make up a closed orbit that system trajectories converge toward as time increases (see [Figure 11.5](#), middle). Limit cycles are paths that the system revisits with great regularity – such that, after a time, the system's behavior is fixed along the path (it is limited within that cycle). This kind of attractor is characteristic of periodic behaviors that exhibit the same stable spatial-temporal pattern over time (e.g., cycle with a stable frequency and amplitude over time) and, moreover, return to the same stable spatial-temporal pattern when perturbed. Rhythmic body and gestural movements provide classic examples and researchers have modeled interlimb (Haken, Kelso & Bunz, 1985) and interpersonal movement dynamics (Schmidt, Carello, & Turvey, 1990), and even complex social behavior, such as speech turns (Buder, [1991](#)), using limit cycle models.

A *strange attractor* is a complex subset of states within a system's phase space that the state variable(s) of a dynamical system evolve toward over time. The term “complex” refers to the fact that strange attractors have a non-integer or *fractal dimension*. That is, the spatial dimension of the subset of states that make up the attractor does not equal the standard Euclidean integer dimensions of one, two or three (e.g., a fractal attractor might live in a two-or three-dimensional state space, but is neither two-dimensional nor three-dimensional, but something in between). Strange attractors are characteristic of chaotic systems (although not exclusively) and in such cases are sometimes referred to as chaotic attractors. The strange attractor displayed on the right of [Figure 11.5](#) is the chaotic attractor of the Lorenz system. The Lorenz attractor is made up of a complex subset of states within a three-dimensional space and has a fractal dimension of approximately 2.05. The set of differential equations that make up the Lorenz system was originally formulated by Edward Lorenz to model atmospheric convection, but it is the strange, butterfly-like structure of its chaotic attractor that has made it so famous.

Order and Control Parameters

Most of the models we have discussed so far have involved state variables that represent an individual behavioral quantity or element. For many complex systems, however, it would be impossible or unfeasible to have a different state variable and equation for every element, agent, or behavioral quantity entailed by the system. For example, if one were trying to model the behavior of gas molecules within a sealed container, it would be impossible to have a set of equations specifying the position and velocity of every molecule. Likewise, if one were attempting to model the movement synchronization that occurs between the members of an audience at a rock concert, it would be impractical to describe the position and velocity of each individual's postural and gestural movements over time. In these and other much simpler cases, it is often better to devise a dynamical model that effectively describes the macroscopic behavior of the system as a whole. This involves identifying a state variable that is able to capture the global or collective organization of the system. Such variables are referred to as *collective variables* or *order parameters*.

Identifying an appropriate order parameter is not always easy and often requires a significant amount of theoretical and empirical work, as well as model testing and assessment (for more details see Kelso, 1995; Nowak & Vallacher, 1998). Once an appropriate order parameter is identified, however, modeling and understanding the dynamics of a complex system is typically much simpler. For example, in the work of Vallacher and colleagues reviewed earlier, where the behavioral synchronization of two individuals is modeled using coupled logistic equations (Equation 11.2), the researchers chose a generic variable they called observable-communicative behavior, and conceived of this as a general patterning of behavior of one person relative to another in interaction. This order parameter is simply the intensity of overall behavior during interaction, which the authors suppose can be modeled as the magnitude of a single dynamical system's state variable. In reality, the social agents in this model are surely employing a whole range of observable behaviors, but by distilling these behaviors into one proposed order parameter, the model is not only more tractable and understandable but also more generalizable.

Another common example of an order parameter is *relative phase*, which captures the location of one periodic or rhythmic behavior in its cycle relative to another. For example, two people walking down the street have the swaying of their arms and legs in a relationship of relative phase. Relative phase is an order parameter because it quantifies in a single measure the spatial-temporal

relationship between two periodic or rhythmic behaviors using a single variable. First employed to capture the stable patterns of synchrony that occur between two mechanical oscillators (e.g., coupled pendulum clocks), it has since been employed to describe numerous biological phenomena, including the behavioral synchrony and social coordination that occurs between the bodily movements of two interacting individuals (see Schmidt & Richardson, 2008 for a review).

A *control parameter* is a system parameter that, when changed, can significantly influence the dynamical regime exhibited by a system. These parameters typically represent some external force, condition, or factor that plays an important role in constraining how a system can evolve over time, and sometimes even the attractors of a system. This is in contrast to other system parameters that are typically fixed and represent non-changing constraints or forces. In the discussion of how an individual's attitude toward group X might be modeled using Equation 11.4, parameter k operated as a control parameter. k represented the amount of positive or negative experiences or information an individual had about group X and increasing or decreasing k influenced attitude strength, with the attitude of individuals toward group X eventually transitioning from negative to positive or vice versa as k was scaled past some critical value.

Bifurcations

Bifurcations are changes in the number and/or type of fixed points or attractors that constrain a system's time evolution and take place when a system's control parameter reaches a particular critical value. Many of the systems described earlier exhibit bifurcations. For instance, the logistic map (Equation 11.2) exhibits a series of bifurcations as the parameter r is increased from 0 to 4. The differential system defined by Equation 11.4 has two bifurcation points, one when $k = +.35$ and one at $k = -.35$. In fact, nearly every system described previously exhibits one or more bifurcations as the values of certain control parameters are increased or decreased.

The possible attractors of a dynamical system that emerge or are destroyed as a control parameter is scaled can be visualized using a bifurcation diagram. The bifurcation diagram for the logistic map (Equation 11.2) as a function of the parameter r scaled from 2.5 to 4 is displayed in Figure 11.6. The y-axis corresponds to the possible long-term values of $x_{(t)}$, with r plotted along the x-axis. It is easy to see from this diagram that the logistic map has a single fixed point for $r < 3$, with the long-term behavior of $x_{(t)}$ equaling a single value of x . At $r = 3$, however, a bifurcation occurs with the possible long-term behavior

equaling two fixed point values of x . The number of fixed points then continues to fork or double as r is further increased (e.g., from 2 to 4, to 8 to 16...fixed points) and eventually exhibits chaotic behavior, with the possible long-term behavior of $x_{(t)}$ equaling all possible values of x .

There are numerous types of bifurcations and not all of them need to be described here (for more information about the different types of bifurcations, see Strogatz, 1994). One theoretical class of bifurcations that are important to mention here are known as *catastrophes*. They are called so because they reflect a qualitatively dramatic change in the behavior of a system. What is particularly interesting about catastrophes and the models used to describe them is that they have proven to be theoretically useful for understanding the sudden emergence and or destruction of a range of social phenomena, including attitudes, self-evaluation, conformity, social relationships, and other catastrophic social transitions (e.g., Guastello, 1995; Latane & Nowak, 1994; van der Mass, Kolstein, & van der Pligt, 2003; Vallacher & Nowak, 1998).

Many of the catastrophic social transitions just mentioned have been modeled using the *cusp catastrophe*. The cusp model is a two-parameter version of Equation 11.4, namely

$$\dot{x} = a + bx - x^3 \quad (11.5)$$

where x is the state variable and a and b are systems parameters. The benefit of this model over Equation 11.4 is that it can be used to model how the interaction of two different external forces can influence behavior. For instance, this model can be used to describe the sudden changes in the dating behavior of couples given two interacting forces: a = love and b = social pressure (Tesser, 1980; Tesser & Achee, 1994). Here, social pressure refers to any family or societal pressure on an individual “not to date” a certain type of individual or group of individuals. For instance, an individual who has a conservative upbringing and whose friends and family are socially and politically conservative might be pressured not to date an individual who has had a very liberal upbringing and whose friends and family are all very liberal. Thus, social pressure is known as a splitting factor or parameter in Equation 11.5.

The bifurcation diagram for Equation 11.5 is displayed in Figure 11.7. The diagram is three-dimensional, with dating behavior, x , on the vertical axis and the control parameters for love, a , and social pressure, b , defining a control surface on the horizontal plane. The folded manifold plotted across the three axis

is a manifold of the fixed points that exist for different settings of a and b . That is, each point on the manifold represents a fixed point, with the points on the non-folded (light grey) area of the manifold corresponding to stable fixed points and the points on the folded (dark) area corresponding to unstable (repulsive) fixed points. A close inspection of [Figure 11.7](#) reveals when the model predicts that a catastrophic change in the dating behavior of a coupled would or would not occur. With respect to the latter case, in which social pressure, b , is low or zero, the prediction (and expectation) is that dating behavior will change almost linearly with changes in love. The more a couple loves each other – the higher the value of a – the more likely a couple is to date. If social pressure is high, however, increases or decreases in love have little effect on the likelihood that a couple will date, unless love increases or decreases past some critical value. At this point the couple will either suddenly enter into a strong dating relationship – if love is increased above a critical value – or suddenly stop dating altogether – if love is decreased below a critical value.

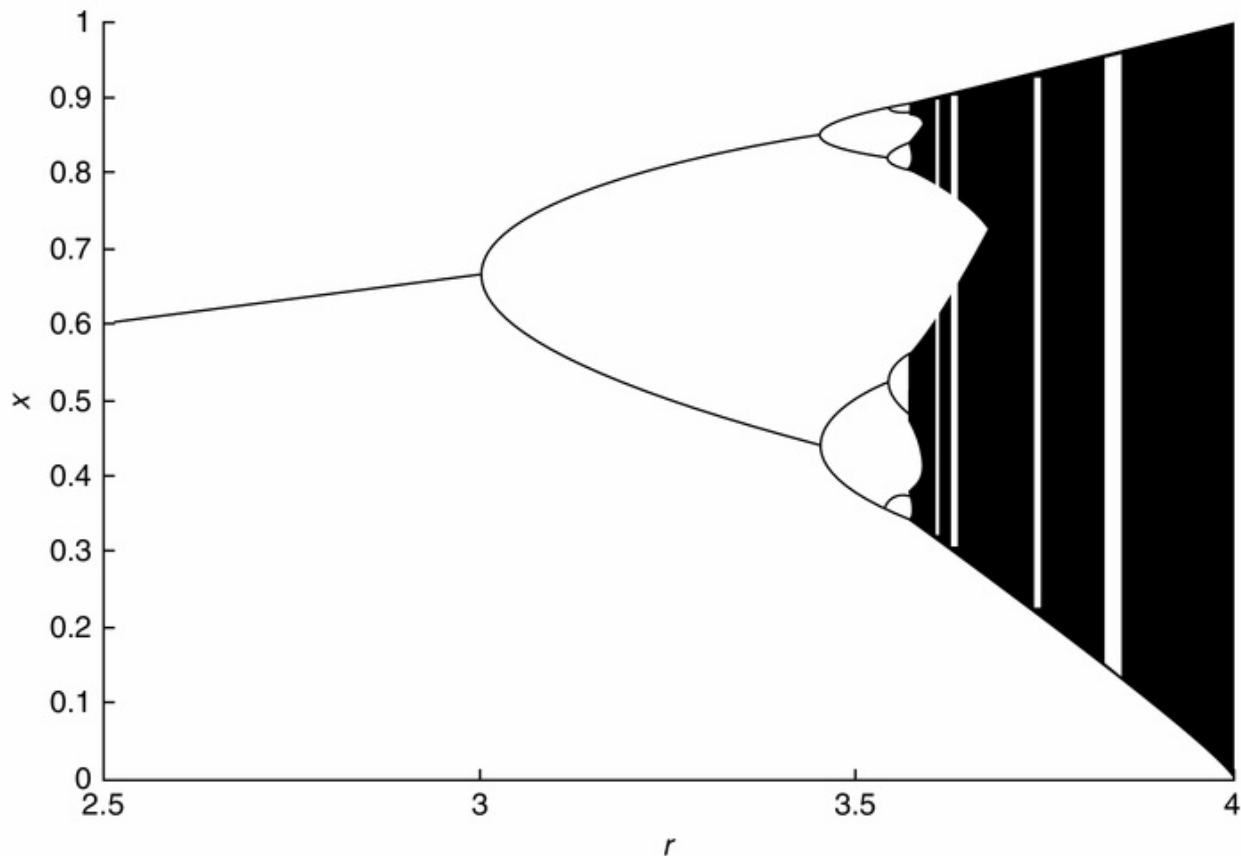


Figure 11.6. Bifurcation diagram for the logistic map, Equation 11.1.

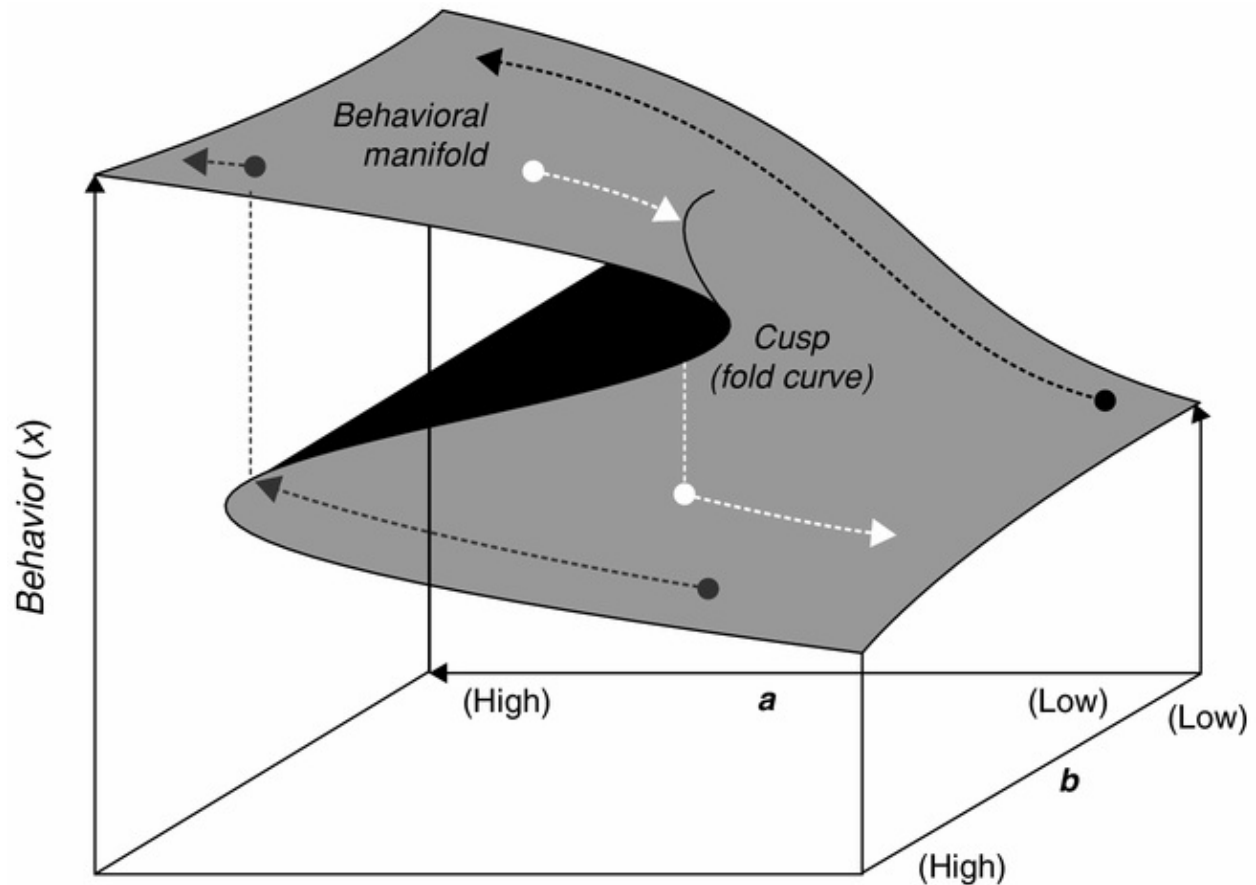


Figure 11.7. Behavioral manifold for the cusp catastrophe. Each point on the grey surface corresponds to the stable fixed points (behavioral attractors) for different a and b values. The dashed lines illustrate the different types of behavioral transitions that are possible when b is fixed at some value and a is scaled up or down.

Cellular Automata, Agent-Based, and Artificial Neural Network Models

Modeling complex nonlinear dynamical systems using difference and differential equations is a powerful way of investigating the stability, dynamic patterns, and self-organization of systems. Using simple deterministic rules, with well-conceived variables and well-grounded parameters, can lead to a meaningful and generalizable theoretical understanding of the dynamics of various types of social behavior. There is, however, another set of methods for modeling the self-organizing dynamics of complex social systems, which starts from a microscopic level, focusing on how phenomena emerge from interacting elements or agents that behave according to simple rules. Such models are

cellular automata and agent-based models.

A *cellular automaton* is a collection of individual cells that form an n -dimensional grid of a specified size and shape, although they are typically restricted to either a one-dimensional line (array) or a two-dimensional square lattice of cells.² Although these cells could represent any component or element, in many social psychological applications a cell represents a person. The cells that make up a cellular automaton represent individual elements of the collective system (e.g., a community of individuals), with the behavior or state of each cell at any point in time determined by a set of simple rules that define its state based on the state of neighboring cells. In a two-dimensional cellular automaton, for example, a cell might be influenced by the four cells directly adjacent to its location (i.e., the cells above and below and to the left and right) or it might be influenced by all eight cells that surround it (i.e., the previous four, plus those on the diagonal).

Cellular automata, including one-dimensional automata, can exhibit a wide variety of complex behavioral patterns with emergent properties characteristic of complex dynamical systems (Wolfram, 2002). The rules that produce such patterns are usually deceptively simple given the complexity of consequences that can follow from them. Within social psychology, cellular automata have been used most notably to model the effects of social influence and public-opinion change based on Latané's social impact theory (for reviews, see Nowak & Lewenstein, 1996; Nowak & Vallacher, 1998). In these models, a two-dimensional cellular automaton is used with cells representing different individuals who possess an attitude of a certain strength toward a topic. The dynamics of an individual's attitudes are determined as a function of the attitude of near neighbors. After a number of iterations, this model eventually results in a stable pattern of attitudes across individuals (i.e., cells). Of particular interest is that the time-evolving patterns that result from this cellular automaton illustrate how key hypotheses from social impact theory play out, such as how pockets of minority opinion emerge over time.

Consider an example. In the models presented by Nowak and Lewenstein (1996), local interaction among cells of a cellular automaton can render small pockets or “walls” of resistance, where a cluster of cells is mutually supportive in sustaining their opinion, despite the onslaught of surrounding opinion. In additional simulations, they show that such minority opinion can grow to become the dominant one. For this to happen, the minority opinion is sparsely distributed in a social environment – weak and scattered. Yet, with a small bias

to favor that minority opinion, things quickly change. What emerges in this model are “clusters” or “bubbles” of minority influence that, as they grow, come to interconnect with each other, thus growing in force and slowly dominating the once-majority opinion. Importantly, all of these simulations depend on simple, local interactions among cells, stretching across space and time.

An extension of classic cellular automata just described, which involve a collection of fixed cells whose state changes over time, is to have cells represent elemental locations, with the state of a cell corresponding to whether it is occupied or not. One of the earliest applications of cellular automata in the social sciences (Shelling, 1971) used this form of cellular automata and demonstrated how social segregation can result from individuals who are more dissimilar to those around them moving to a different location.

Agent-based modeling is another extension of cellular automata models. Such models have more potential for complexity, as agents in these models are not confined to a matrix of cells but are able to move around. One prominent use of agent-based modeling comes from the work of Axelrod (1984; Axelrod & Dion, 1988; Axelrod & Hamilton, 1981) on the emergence of cooperative behavior in games like the prisoner's dilemma. Axelrod's work demonstrates that strategies that are cooperative but punish defections (e.g., the tit-for-tat strategy) are most successful in computer simulation tournaments in the long run when pitted against agents using other strategies. Moreover, if the program increases the likelihood of an agent taking on the strategy of an interactant who is more successful in their game, the tit-for-tat strategy spreads in a population. Intriguingly for a dynamical perspective, history is crucial for learning such a strategy. If there is a reshuffling of agents after each trial so that they have no continuity in whom they interact with, then there is no means for an agent to learn personally the negative consequences of defecting when the other cooperates (punishment on the next trial) – the agents have by then moved on to the next partner. As a result, lacking history with an interactant cooperative strategies are not learned (Axelrod, Riolo, & Cohen, 2002) – the tit-for-tat strategy does not spread.

One interesting aspect of cellular automaton and agent-based models more broadly is the emergent phenomena that result from the structure of linkages between agents. Traditional social psychological methods do not account well for how being in contact with multiple individuals, repeatedly and over time, affects behavior (Mason, Conrey, & Smith, 2007). Because agent-based models do, they allow a researcher to see how novel phenomena may emerge as a

consequence of the interdependencies among agents (Smith & Conrey, 2007). A recent agent-based simulation of impression formation capitalizes on the unique potential for such models to examine the consequences of flow of information across different types of linkages (Smith & Collins, 2009). In Smith and Collins's model, participants can sample information about another person directly or indirectly (e.g., through gossip). They used Kenny's Social Relations Model (1994) to analyze key features of impressions – for example, the degree to which impressions do in fact reflect commonalities (across perceivers) attributable to the target, versus how much variance is owing to perceiver and specific target-perceiver relationships. The results of their simulations indicate that the ways of obtaining information matter. The most negative impression came from one-sided elicitation (because people are likely to cease seeking information if their initial impression is negative). Sampling information socially led to more positive impressions on average, as well as reduced perceiver effects and relationship effects. Moreover, the authors found emergent phenomena that had to do with dyadic reciprocity and the relative accuracy of generalized versus dyadic accuracy (Smith & Collins, 2009).

In the Smith and Collins's (2009) model, although the model is directed toward understanding cognitive processes through transmission of information about a person (directly or third-person), the interacting agents in the model are individuals. It is important to realize, however, that the agents in agent-based modeling can be the cognitive elements themselves (e.g., the interaction between visual perception, inferences, judgments, etc.). Thus, there is a close link between such models and dynamic models of memory. A brand of modeling that has focused entirely on exploration of the dynamics of social cognitive processes is *artificial neural network models* (also called connectionist models in the late 1990s). In a classic work in social psychology, Smith and DeCoster (1998) used a basic associationist neural network model to explore a wide range of phenomena, including stereotyping and person perception. They argued that neural network models of memory and social judgment – which function by integration of basic informational cues (much like interaction among agents) – can more parsimoniously account for a range of social phenomena than traditional models of memory (e.g., Wyer & Srull, 1989). Modern neural network models can also be used to model cognitive dynamics, including the dynamics of social cognition; see Smith (1996) for an excellent early introduction to use of neural network (connectionist) models in social psychology.

Such models have also been applied to social interactive phenomena. Nowak and Vallacher (1998) review early models of interpersonal dynamics using

neural network models. Perhaps the best illustration of such models is a very recent model meant to tap into the fast-time-scale dynamics of person perception and judgment by Freeman and Ambady (2011). They proposed that person construal (such as identifying the gender of a person) is constrained by a constellation of information sources in the environment, such as hair cues, skin cues, facial configural cues, and so on. They developed an interaction-activation framework (see collection of papers in Rumelhart & McClelland, 1985) in which cues are integrated *incrementally* and *probabilistically* in time. This provides an array of predictions about how personal construal *dynamics* is shaped and guided by different combinations of cues. The authors have explored such dynamics in a variety of experiments on the behavioral dynamics of these perceptions using, for example, mouse-tracking methods. In this behavioral approach, researchers can track participants as they move their computer mouse toward varying options on a screen (e.g., during person construal). These mouse movement trajectories can then be mapped directly onto a computational neural network model: The evolution of the computer cursor on the screen toward a response box can be captured by a neural network state, achieving some stable response activation over choices. Accordingly, the dynamical model, together with behavioral recordings, makes a tidy dynamical package for exploring the dynamics of social decision making, perception, and judgment. For a review of the behavioral results, see Freeman *et al.* (2011); for software that allows dynamic tracking of these social processes, see MouseTracker (Freeman & Ambady, 2010).

Dynamical Systems Analysis

Dynamical modeling is important because it provides researchers with a set of tools for understanding in an abstract and often extremely general way how human and social behavior can emerge. As many of the examples provided earlier highlight, the power of a dynamical model does not always rest on its ability to simulate real-world behavior, but rather whether it can generate testable predictions, enhance theoretical development, and motivate research questions. Unfortunately, there is no step-by-step guide that one can follow when developing dynamical models of human and social behavior. Building an effective model requires good understanding of the many different types of models and mathematical functions that can be employed to capture differing types of dynamics, as well as a significant amount of theorizing and lots of trial and error. A researcher interested in modeling the dynamics of a behavioral

system also needs to have a good understanding of the system's underlying stabilities and its relevant state variables and parameters. In many instances, however, researchers in social-personality psychology do not start with sets of state variables, parameters, or mathematical functions or equations, and may not know the nature of a behavioral system's underlying dynamics. In such cases, research typically starts with a temporal sequence of behavioral measurements or observations – a *behavioral time series* – recorded during experimental, nonexperimental, or observational research. A researcher then attempts to uncover the dynamics of a behavior using various forms of time-series analysis. Accordingly, in this final section of the chapter we review some of the tools that can be employed for dynamical analysis of behavioral time series.

Before introducing various methods of dynamical time-series analysis, it is important to appreciate that empirical research and the analysis of behavioral time-series data can be, but is not always, a precursor to modeling. Rather, research and modeling are best conceptualized as complementary methods of dynamical analysis, with researchers often moving back and forth between both forms of research (i.e., behavioral research and dynamical modeling), using experimentation and time-series analysis to identify key state variables, attractors states, and control parameters, and mathematical modeling to better understand and test empirical findings and make future predictions. A detailed description of the nuances of how one goes from the dynamical analysis of behavioral time-series data to a dynamical model is well beyond the scope of this chapter. We do wish to emphasize, however, that building a dynamical model is not always necessary for understanding the dynamics of behavior. In many cases, building a model to simulate the dynamics uncovered via behavioral time-series analysis will not necessarily provide more insights or additional information about a system's underlying dynamics. Accordingly, many dynamical systems researchers are less concerned with building dynamical models and instead focus more of their efforts on uncovering the dynamics of behavioral systems via experimentation and the kinds of dynamical time-series analysis techniques outlined in the following sections.³

Behavioral Measurement

As with any research study, determining valid and reliable dependent variables is fundamental. What the right dependent variable is when investigating the dynamics of social phenomena will of course depend on what behavior is measurable in a given context, along with a researcher's theoretical interests.

Obviously, the dependent variable should capture the *state* of the agent or system at the time of measurement. As we outlined previously, a behavioral state represents a wide array of measures. For example, socially relevant measures may come in the form of self-esteem, personality characteristics, attitude, or social dominance measured over time. A measured behavioral state could also be a mode or type of behavior, such as whether an individual carries an object alone or together with another person, or exclusionary acts of a pair of individuals with respect to some third (perhaps out-group) individual during an interactional game. In social-personality psychology, it is also common for researchers to record other more indirect or implicit measures of a behavioral state, including physiological measures such as heart rate, cortisol level (Dickerson & Kemeny, 2004), muscle activity, skin conductance, and neurophysiological measurement techniques (e.g., using EMG, EEG, or fMRI; see Berkman, Cunningham, & Lieberman, Chapter 7 in this volume; Cacioppo, Tassinary, & Berntson, 2007; Stam, 2005). Overt behaviors that have social relevance also include body position and movement (for reviews, see Fowler, Richardson, Marsh, & Shockley, 2008; Schmidt & Richardson, 2008) and eye gaze (Richardson & Dale, 2005). Any social process likely has a behavioral proxy that can be tracked, semi-continuously, in time.

No matter what dependent variable one chooses, there are several important requirements when investigating the dynamics of social behavior. First, and most obviously, the dependent variable must correspond to a behavioral state measurement that can be recorded repeatedly over time. This results in a sequence of measurements over time, or more specifically, a *behavioral time series*. Here, the term “time” could refer to “clock” time such as second, minute, hour, day, month, etc., or it could refer to some other time scale, such as trials, sessions, or events. That is, a researcher could record a behavior almost continuously, making measurements several times a second or after some longer time interval. For instance, a researcher could record an individual's postural position 50 times a second during the course of a 2-minute conversation (e.g., Schmidt, Fitzpatrick, Caron, & Mergeche, 2011; Shockley, Santana, & Folwer, 2003), the heart rates of an infant and mother sampled 1,000 times a second during a three-minute face-to-face interaction (Feldman, Magori-Cohen, Galili, Singer, & Louzoun, 2011), vocal activity assessed at fractions of a second (Warlaumont, Oller, Dale, Richards, Gilkerson, & Dongxin, 2010), respiration patterns assessed at each breath intake for dyads involved in lengthy casual conversations (McGarva & Warner, 2003; Warner, 1992; Warner, Waggener, & Kronauer, 1983), an individual's emotional expression twice a minute while

watching an emotive film (e.g., Mauss, Levenson, McCarter, Wilhelm, & Gross, 2005), or an individual's self-esteem or mood every day over the course of two years (e.g., Delignières, Fortes, & Ninot, 2004).

Behavioral time series can also be sequences of discrete behavioral events (e.g., occurrence of categorical events or coded behaviors) that could be dichotomous (a single action occurs or not, such that the time series is a sequence of 1s and 0s) or that involve several different discrete states recorded on a nominal unit scale. For instance, a researcher could record which object is being looked at and in what sequence during a joint task (Richardson & Dale, 2005), which words and sequences of words are used by an individual or group of individuals during a conversation (Louwerse, Dale, Bard, & Jeuniaux, 2012), or a dichotomous coding of when individuals vocalize or not during interactions with someone they believe to be attitudinally similar or dissimilar to themselves (McGarva & Warner, 2003).

Irrespective of the type or time scale of the behavior being measured or recorded, the ordering of observations or measurements in the behavioral time series must be recorded sequentially in time. Extracting the emerging patterns or stable states of behavior from data requires historical information about the state of the system preceding its current state. If future observations depend on observations that preceded it in time – in other words, that are sequentially dependent – then data recorded nonsequentially will prevent identification of trends, stable states, or reoccurring patterns that may exist. In many dynamic analyses, an additional constraint is that the time intervals between sequential measurements must be the same. This is because the dynamic regimes that characterize many continuous behaviors have specific temporal properties that can only be determined if the time between measurements is known and equivalent.⁴ For instance, many human and social behaviors are characterized by periodic patterns, from the intrinsic periodicity of brain-wave patterns (e.g., Tognoli, Lagarde, DeGuzman, & Kelso, 2007), to the leg and arm movements of an individual while walking (e.g., Moussaïd, Perozo, Garnier, Helbing, & Theraulaz, 2010), to the day-to-day mood of an entire population of people (Dodds, Harris, Kloumann, Bliss, & Danforth, 2011). In each case, the same or similar behavioral states recur again and again after a certain period of time (i.e., with a certain temporal frequency).

It is also crucial that the measurement device has enough resolution to reliably capture whether the behavioral state has changed across repeated measurements. In many instances, traditional methods of measurement, such as obtaining self-

evaluations on a 7-or 9-point Likert scale, or simply coding whether or not a certain behavior or action occurred (i.e., 1 for yes and 0 for no), can provide the resolution needed. In other cases, such as when recording the subtle gestures, postures, or eye movements of conversing individuals, or the subtle changes in the body language, emotion, and anxiety of an individual during an interpersonal confrontation, more advanced methods of measurement might be required. Thankfully, recent technical progress has facilitated collection of such behavioral time series. For instance, modern video processing technology enables researchers to acquire objective whole-body activity time series of one or more individuals from synchronized multiview video recordings (e.g., Kupper et al, 2010; Schmidt, Morr, Fitzpatrick, & Richardson, 2012). There are also technologies for continuously recording an individual's movements in space and time, such as Polhemus tracking systems (e.g., Polhemus Liberty and Latus systems, Polhemus, Ltd, Virginiaia) and NDI Optorack (Northern Digital Inc., Ontario, Canada). There are also now a number of low-cost off-the-shelf gaming systems (e.g., Nintendo Wii remotes and force plates, and the MS Kinect) that can be used to wirelessly record the movements of interacting individuals in real time. As for physiological measures, Biopac systems are now widely used for tracking one or more signals, and can be used as both an electroencephalogram and electromyogram (Biopac Systems, Inc., Aero Camino Goleta, CA). Blascovich (Chapter 6 in this volume) provides more details about physiological measurement.

With regard to recording more discrete behavior, even for eye movements and gestures, as well as language analysis, there are now hardware and software applications that can automatically categorize the behaviors being emitted. For language analysis, researchers relied on well-developed schemes in the psychological sciences for coding or transcribing social interaction, involving time-intensive practices that require careful selection of units of analysis, guided by research goals (Bakeman, Deckman, & Quera, 2005; Heyman, Lorber, Eddy, & West, Chapter 14 in this volume; Kreuz & Riordan, 2011). Such research can be facilitated by powerful annotation software (Loehr & Harper, 2003). Eye movement technologies can use “areas of interest” (AOI) to transform a sequence of x,y-coordinates on a computer screen to a set of looked-to objects. Researchers have also used continuously recorded body movements and speech to infer discretely labeled states. For example, in the domain of human-computer interaction, machine-learning algorithms have been applied to extract meaningful states (Castellano, Kessous, & Caridakis, 2008), such as which emotion is being experienced by a person, given multimodal cues such as

speech, and face and body movements. In addition, gesture recognition algorithms allow hand gesture patterns (discrete categories) to be obtained by learning algorithms applied to body movement data (see Mitra & Acharya, 2007 for a review).

Methods of Dynamical Analysis

So what do you do after you have obtained a time-series recording of a behavioral phenomenon? How do you investigate the dynamics present in a recorded time series? In general, the dynamical analysis of a behavioral time series involves qualitative and graphical assessment of the time-evolving pattern of behavior and then quantitative linear and/or nonlinear time-series analysis.

Qualitative and graphical assessment. For any research study, knowing what the recorded data “look” like is essential for appropriate understanding and analysis. Although visual inspection alone does not typically reveal what the underlying dynamics are, it does provide a general understanding of the kinds of analysis techniques that will be needed to uncover the dynamics. For dynamical analysis, this first and foremost involves graphing the behavioral time series on a time-series plot and visually inspecting the patterns it contains. For illustrative purposes, hypothetical examples of some of the different kinds of time series that might be obtained in social-personality psychology are displayed in [Figure 11.8](#) (consult sources cited later in this paragraph for data examples). An inspection of different time series highlights just a few of the many different types of behavioral time-series patterns that could be recorded. In some cases the patterns of change over time are relatively simple and regular: the monotonic decrease of an individual's anxiety level over the course of 50 therapy session (Heath, 2000) and the oscillatory movements of an individual's right arm while walking (Harrison & Richardson, 2009). In other cases the patterns of change over time are highly complex and appear to be nondeterministic or stochastic (i.e., random): an individual's self-esteem over the course of 1.5 years (see Delignières et al., 2004) and the trial-by-trial RT of an individual completing a 512 trial lexical decision task (see Holden, 2005). Others seem to fall somewhere in between, containing semi-periodic patterns or other complex regularities. Two examples are the daily hedonic level or mood of individuals over the course of twelve weeks (see Larsen & Kasimatis, 1990) and the eye fixations that occur when an individual scans the world during a conversation (see Richardson & Dale, 2005).

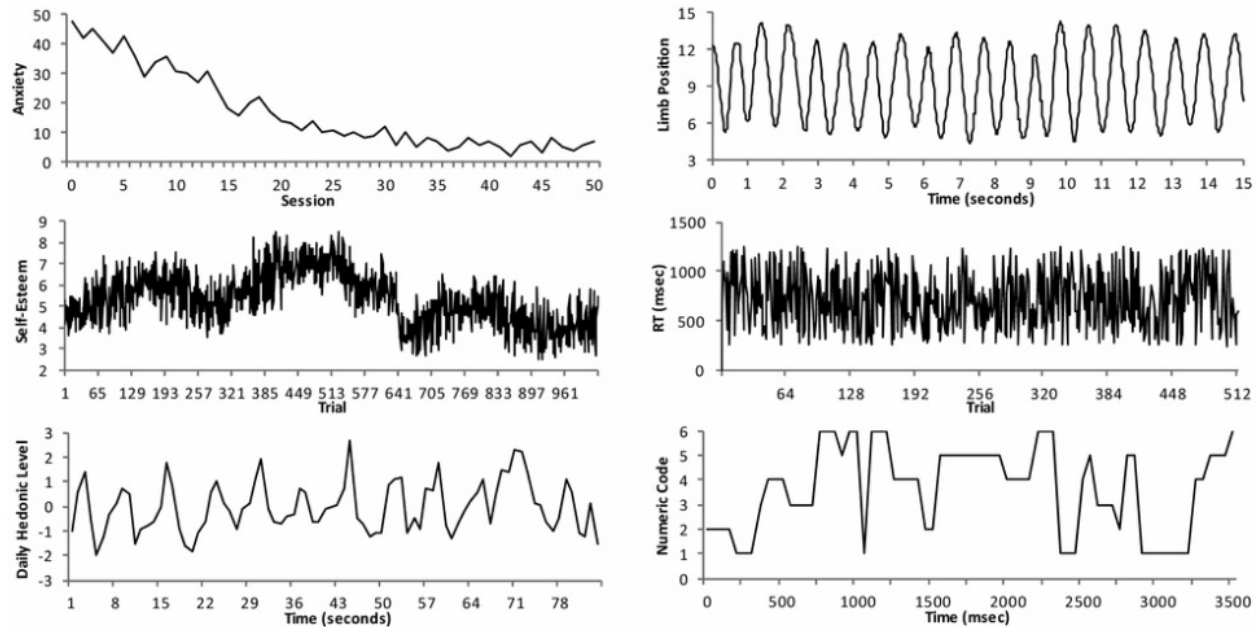


Figure 11.8. Hypothetical examples of several types of behavioral time series. (top left) Change in anxiety level for an individual over 50 therapy sessions. (middle left) An individual's self-esteem recorded on a 9-point Likert-scale twice a day for 512 days. (bottom left) An individual's daily hedonic (mood) level recorded over 12 weeks. (top right) Motion sensor recording of an individual's right arm movements while walking. (middle right) Reaction times of a participant completing a 512 trial lexical decision task. (bottom right) A time series representing categorical data obtained from eye movement behavior while a person views an array of 6 images. An eye-tracking system records a numeric identifier 1–6 reflecting which particular image is being fixated over time (see text for more details).

In addition to inspecting time-series plots of one's data, plotting a behavioral time series as a trajectory in phase space can also provide researchers with a clear qualitative understanding of the attractor(s) that constrain the time-evolving behavior. If the state space of a behavioral systems is known a priori (and contains no more than three dimensions), this is often quite easy. It is more often the case, however, that the phase space of a behavioral system is not known a priori. In such cases, the process of plotting behavioral time series and a trajectory in phase space requires that one uncover – or more precisely, recover – the phase space of a behavioral system analytically. *Phase space reconstruction* involves a number of steps that enable a researcher to recover a phase space isomorphic to the system's real phase space. In short, phase space reconstruction involves extracting the entire multidimensional dynamics of a system, in all its

complexity, from a one-dimensional time-series recording. Although a detailed description of the steps required to complete phase space reconstruction is too involved to be unpacked here (for a more detailed description and tutorial, see Abarbanel, 1996 and Kantz & Schreiber, 1997), the method itself is powerfully intuitive. For systems that have a phase space with more than three dimensions, phase space reconstruction also provides a quantifiable measure of a system's dimension – an indication of the number of the state variables required to model the system effectively. Examples of phase space reconstruction applied to a socially relevant phenomenon are described in Shockley *et al.* (2003) and Richardson, Schmidt and Kay (2007), who study behavioral synchrony between interacting partners. In this research, a single bodily movement measure – such as postural change – is recorded as a signature of the fluctuations of the total mind/body system. To gain access to the higher dimensions of the system from the (single) one-dimensional time-series measure (i.e., change in posture sway over time), researchers use phase space reconstruction to discern how many dimensions best capture the fluctuations observed in the movement time series (often as many as 10 dimensions). Once phase space is reconstructed, researchers can then mathematically compare how two people's movements change in relation to each other in this higher-dimensional space.

Linear and nonlinear time-series analysis. One of the key decisions a researcher must make when inspecting time-series plots is whether the dynamic regime that characterizes the behavior of interest is simple or regular enough to be analyzed using linear methods, or whether the behavioral dynamics are sufficiently complex that one must employ nonlinear methods. Unfortunately, there is no definitive rule as to when one should employ linear or nonlinear methods and in many instances, especially when performing a dynamical analysis on new phenomena or on behavioral time series that have not previously been examined, it is prudent to employ a range of linear and nonlinear methods in order to determine different aspects of the behavioral dynamics recorded.

In general terms, however, linear methods of analysis are preferable when the patterning of movement or behavior being investigated is highly regular and *stationary*. That is, the mean and dispersion of sampled values in the time series have a regular pattern and remain more or less the same across the interval of recorded time. The anxiety, daily hedonic level, and limb movement time series in Figure 11.8 all meet this criteria, as does the RT time series (although see section on fractal analysis later in the chapter). For time series data that is *nonstationary* – the mean and dispersion of sampled values vary markedly across the time-series recording – or for behavioral time-series that contain a

high degree of stochastic variability or involve highly complex or aperiodic patterns of change over time, nonlinear methods may be more effective. What follows is a brief description of several common and generally applicable linear and nonlinear time-series methods for research in social-personality psychology (for more detailed discussion and tutorials, see Abarbanel, 1996; Boker & Wenger, 2007; Gottman, 1981; Heath, 2000; Kantz & Schreiber 1997; Riley & Van Orden, 2005).

Spectral analysis and cross-spectral coherence. One of the first questions commonly asked when analyzing time-series data is whether the data contains any periodic or temporal structure. Consider the limb movement and daily hedonic time series in Figure 11.8. Do the up and down fluctuations occur in a stable periodic manner? If so, after what period of time (i.e., at what frequency) does the pattern repeat itself? Conducting a spectral analysis enables one to answer these questions by decomposing a time series into its periodic components by estimating how well a set of sine or cosine functions of different frequencies and amplitudes fit the data. Performing a spectral analysis is much like conducting a regression analysis in that you are attempting to decompose the major sources of variation in the data, in this case trying to determine which component frequencies account for significant amounts of variability in the signal. For highly stable periodic behavior, like the rhythmic limb movements displayed in Figure 11.8, there is usually only one dominant or fundamental frequency component. For less stable periodic data, like the daily hedonic time series displayed in Figure 11.8, individual frequency components will be less powerful. There might also be more than one frequency component in a time series (i.e., multiple frequency components). This is particular true for highly complex or semi-periodic time series.

Spectral analysis can also be employed to determine how correlated two time series are by examining, essentially, the similarity of their frequency patterns. This comparison is called cross-spectral coherence and indexes the correlation between two time series on scale of 0 to 1, and is analogous to calculations of the squared correlation coefficient (Gottman, 1981; Porges et al., 1980; Warner, 1988). In social-personality research, cross-spectral coherence is commonly employed to examine mutual influence and behavioral coordination, that is, the degree to which one individual's behavior is influenced by and/or coordinated with the behavior of another. For example, Sadler *et al.* (2009) used cross-spectral methods to explore the rhythmic relationships between two people's dominance and affiliative dynamics during interaction. In this study, coders dynamically tracked interaction partners using a joystick, thus producing

dominance/affiliation time series. The authors found that, indeed, pairs of interaction partners exhibit similar affiliation amplitude-frequency patterns. Put differently, they shared behavioral cycles.

Autocorrelation and cross-correlation. Dynamic human and social behavior is usually correlated over time. In other words, how an individual behaves at any given moment is typically correlated with how the individual behaved sometime recently. The dependence or correlation between future and past behavior can be determined using autocorrelation. Sometimes called *lagged correlation*, autocorrelation identifies if future states are correlated with past states by determining the correlation between points in a time series at different time lags. A positive autocorrelation indicates persistence of behavior after some time lag; the behavioral change is similar from one observation to the next (e.g., positive changes or state values in both past and future states). A negative autocorrelation indicates anti-persistence of behavior after some time lag; opposite behavioral change occurs from one observation to the next (e.g., positive changes or values in the past state correspond to negative changes or values in the future states).

Cross-correlation is a simple extension of autocorrelation and examines the dependence or correlation between future and past values of different time series. It is also commonly used to examine mutual influence and behavioral coordination, and yields similar results to coherence analysis (described earlier), except that one can also look at the correlation between individuals' behaviors at time lags other than zero. Accordingly, it can be employed to determine if two behaviors are attracted toward each other, and also whether one behavior leads or follows another behavior at some specific time lag.

Relative phase analysis. Another technique for examining mutual influence and behavioral coordination is relative phase analysis. This technique has been employed most extensively in research examining behavioral synchrony – the rhythmic movement coordination that occurs between the limb or body movements of interacting individuals (e.g., Marsh, Richardson, & Schmidt, 2009; Miles, Lumsden, Richardson, & Macrae, 2011; Schmidt & Richardson, 2008). It can also be employed to investigate the patterns of coordination that occur between any set of rhythmic or periodic behavior (e.g., the coordinated changes in day-to-day mood of husband and wife or mother and child). In short, the technique involves calculating the difference in the “phase” of two (or more) rhythmic or periodic behaviors over time (for details on multivariate relative phase analysis, see Frank and Richardson, 2010; Richardson, Garcia, Frank, Gregor, & Marsh, 2012). Here the term “phase” refers to the location of a system

or behavior within its cycle. The relative phase or “difference in phase” between two rhythmic or periodic behaviors therefore corresponds to the location of one behavior within its cycle relative to the location of the other within its cycle. Thus, if the relative phase between two behavioral time series remains the same over time, the behavior is said to be coordinated at that relative phase relation. For example, consider two people coordinating their rhythmic gait while walking down the street together – their leg cycles are in the same place and are cycling together.

Typically, behavioral synchrony is constrained to two stable patterns of behavioral coordination over time, commonly referred to as *inphase* and *antiphase* coordination (Haken, Kelso & Bunz, 1985; Schmidt, Carello, & Turvey, 1990). Inphase coordination corresponds to rhythmic or periodic movements or behaviors that move or change in the same direction at the same time (such as the walkers we just mentioned). Antiphase coordination corresponds to rhythmic or periodic movements or behaviors that move or change in the opposite direction at the same time. To use the walking example again, this would mean that individuals would be continuously moving their legs in opposite patterns – as one person swings her right leg forward, the other would be swinging her right leg back. It is worth noting that these latter descriptions of inphase and antiphase coordination characterize perfect or absolute synchrony. During natural social interaction, however, the movements or behavior individuals do not usually become coordinated in a perfect inphase or antiphase manner, but rather exhibit intermittent periods of inphase and/or antiphase coordination (e.g., Richardson, Schmidt, & Kay, 2007; Schmidt & O’Brien, 1997).

Recurrence analysis. The analysis methods discussed so far are based primarily on assumptions of linear relations that underlie most analyses commonly used in psychology (e.g., ANOVA). Accordingly, they are only able to capture the linear dynamics of stationary time-series data. Recurrence analysis, however, is a nonlinear analysis method and can be employed to analyze both stationary and nonstationary data. Indeed, the beauty of recurrence analysis, in comparison to other linear time-series methods, is that it does not require assumptions about the structure of the time series being investigated or the underlying dynamics that shape the recorded structure: The behavior can be periodic, nonperiodic, or stochastic, even discrete or categorical.

Although recurrence analysis is still relatively new, particularly in psychology (e.g., Riley, Balasubramaniam, & Turvey, 1999; Shockley, Santana, & Fowler,

2003), there is now substantial evidence that suggests it is potentially one of the most robust and generally applicable methods for assessing the dynamics of biological and human behavior (e.g., Marwan & Meinke, 2002; Zbilut, Thomasson, & Webber, 2002), including social behavior (e.g., Dale & Spivey, 2005; Richardson et al., 2008; ; Shockley et al., 2003). Essentially, recurrence analysis identifies the dynamics of a system by discerning (1) whether the states of the system behavior recur over time and, if states are recurrent over time, (2) the degree to which the patterning of recurrences are highly regular or repetitive (i.e., deterministic). Conceptually, performing recurrence analysis on behavioral data is relatively easy to understand; one simply plots whether the recorded points, states, events, or categories in a time series or behavioral trajectory are revisited or reoccur over time on a two-dimensional plot, called a recurrence plot. This plot provides a visualization of the patterns of revisitations in a system's behavioral state space and can be quantified in various ways – a process known as recurrence quantification – in order to identify the structure of the dynamics that exist (see Marwan, 2008 and ; Weber & Zbilut, 2005 for more detailed reviews). The plots in Figure 11.9 are examples of what recurrence plots look like for a categorical (left plot) and continuous (right plot) behavioral time series.

Like spectral analysis and autocorrelation, recurrence analysis can also be extended to uncover the dynamic similarity, mutual influence, or coordinated structure that exists between two different behavioral time series or sequences of behavioral events. This latter form of recurrence analysis is termed *cross-recurrence analysis* and is performed in much the same way as standard (auto) recurrence analysis. The key difference is that recurrent points in a cross-recurrence plot correspond to states, events, or categories in two time series or behavioral trajectories that are recurrent with each other. Cross-recurrence analysis can therefore be employed to capture and then quantify the co-occurring or coordination dynamics of two behavioral time series or discrete behavioral sequences. Accordingly, some researchers have adopted cross-recurrence analysis to investigate semantic similarity in conversation (Angus, Smith, & Wiles, 2011), perceptual-motor synchrony between people interacting (Shockley et al., 2003; Richardson & Dale, 2005), and vocal dynamics during development (Warlaumont et al., 2010).

Fractal analysis. Researchers in social-personality psychology (or in any other field of psychology) commonly collapse repeated measurements into summary variables, such as the mean and standard deviation, under the assumption that the measured data contains uncorrelated variance or random

fluctuations that are normally distributed. With respect to dynamic behavioral time-series data, however, this is rarely true, and thus summary statistics such as the mean and standard deviation often reveal little about how a system evolves over time. Indeed, time-series recordings of human performance and behavior typically contain various levels of correlated variance or data fluctuations (i.e., nonrandom fluctuations) that are not normally distributed (Stephen & Mirman, 2010) and, moreover, are structured in a *fractal* or *self-similar* manner (Gilden, 2001, 2009; Van Orden, Holden, & Turvey, 2003; Van Orden, Kloos, & Wallot, 2011).

A fractal or self-similar pattern is simply a pattern that is composed of copies of itself nested within itself. As a result, the structure looks similar at different scales of observation (i.e., magnification). Conceptually similar to geometric fractal patterns (Mandelbrot, 1982), a fractal time series is therefore a time series that contains nested patterns of variability (see Figure 11.10). That is, the patterns of fluctuation and change over time look similar at different scales of magnification or measurement resolution (i.e., as one zooms in and out).⁵ The self-esteem time series in Figure 11.8 is a good example of a fractal or self-similar time-series pattern. This time series is displayed again in Figure 11.10, with the self-similarity of its temporal fluctuations revealed by zooming in on smaller and smaller sections. At each level of magnification the temporal pattern looks similar (see Bassingthwaite, Liebovitch, & West, 1994 or Holden, 2005 for a more detailed tutorial).

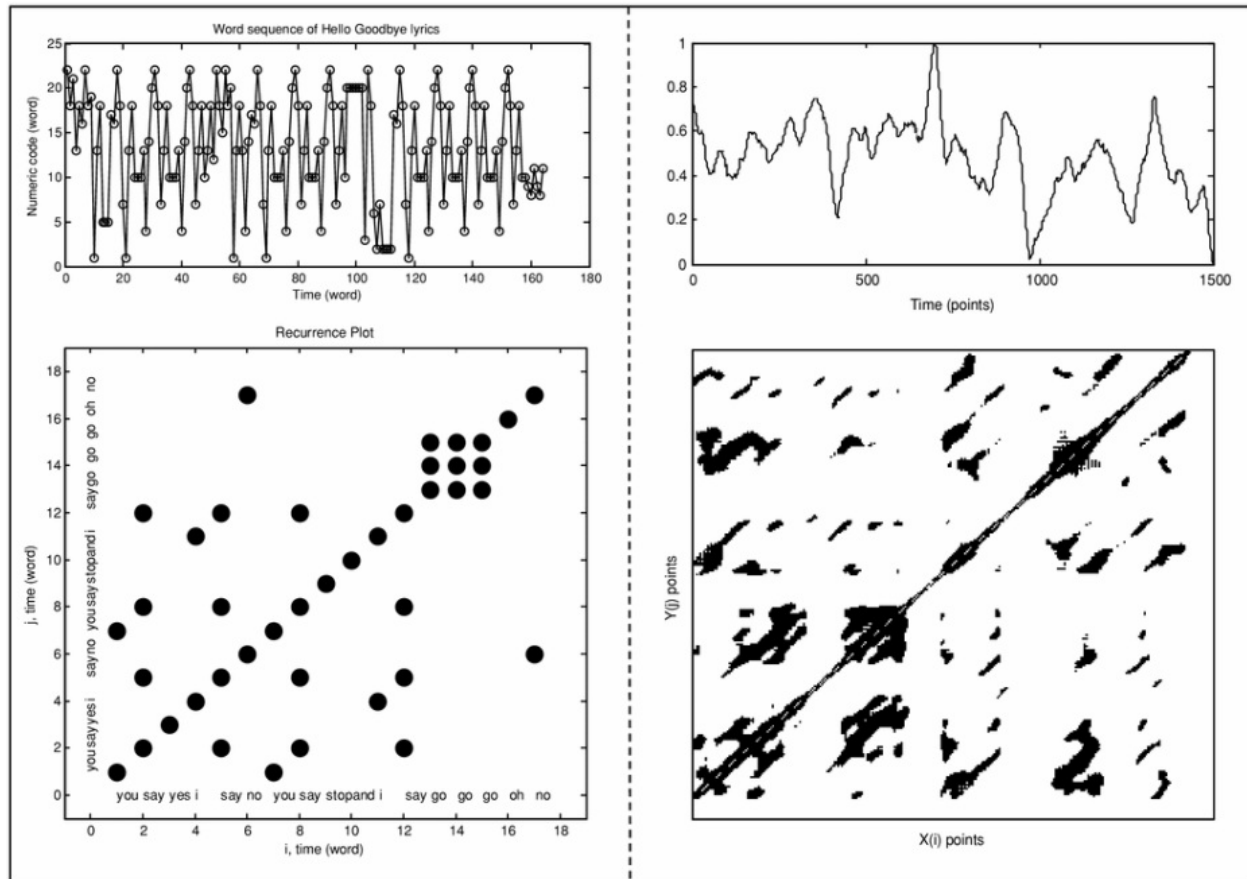


Figure 11.9. (top left) The full time series of words extracted from the lyrics of “Hello, Goodbye” by the Beatles. The y-axis represents the numeric identifier to which a word is assigned, and the x-axis represents word-by-word unfolding of this “lexical” time series. (bottom left) A recurrence plot of the first 20 words in the lyrics. Each point on the plot represent a relative point (i,j) in the lyrics at which a word is recurring. The “go go go” usage appears as a particular “texture” on the plot, along with the two-word sequence “you say” appearing as a diagonal line structure. (top right) The anterior-posterior postural sway movements of single individual standing and listening to another person speak for 30 seconds, recorded at 50 samples a second. (bottom right) A recurrence plot of the first 10 seconds of postural data. Despite the nonperiodic and nonstationary pattern of the postural sway movement, the recurrence plot reveals a significant degree of recurrent and deterministic structure, with the patterns of postural behavior reoccurring over time.

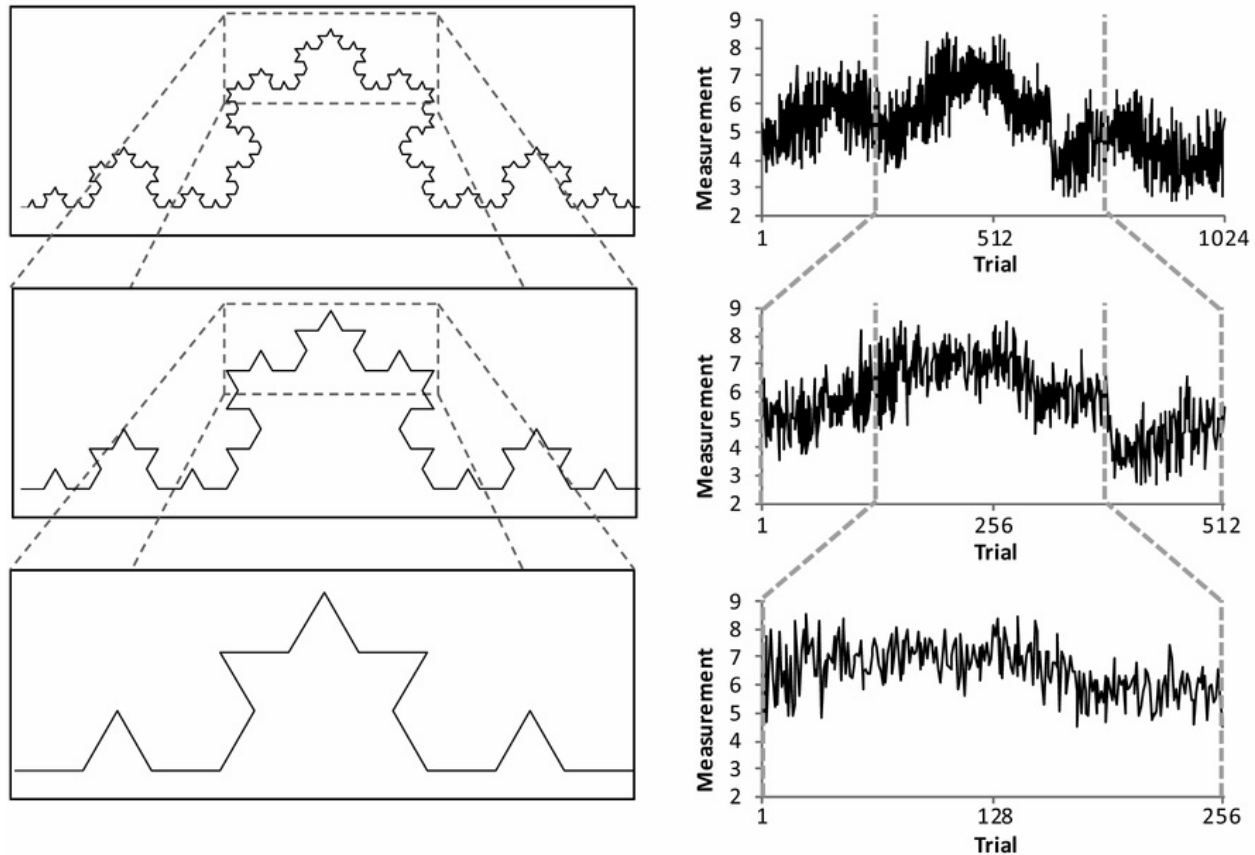


Figure 11.10. Example geometric and temporal fractal patterns. (left) Koch Snowflake at three levels of magnification. (right) The repeated self-esteem measurements presented in [Figure 11.9](#) at three levels of magnification. The factual nature of these patterns is revealed by self-similar patterns being observed at smaller and larger magnitudes of observation (adapted from Holden, 2005).

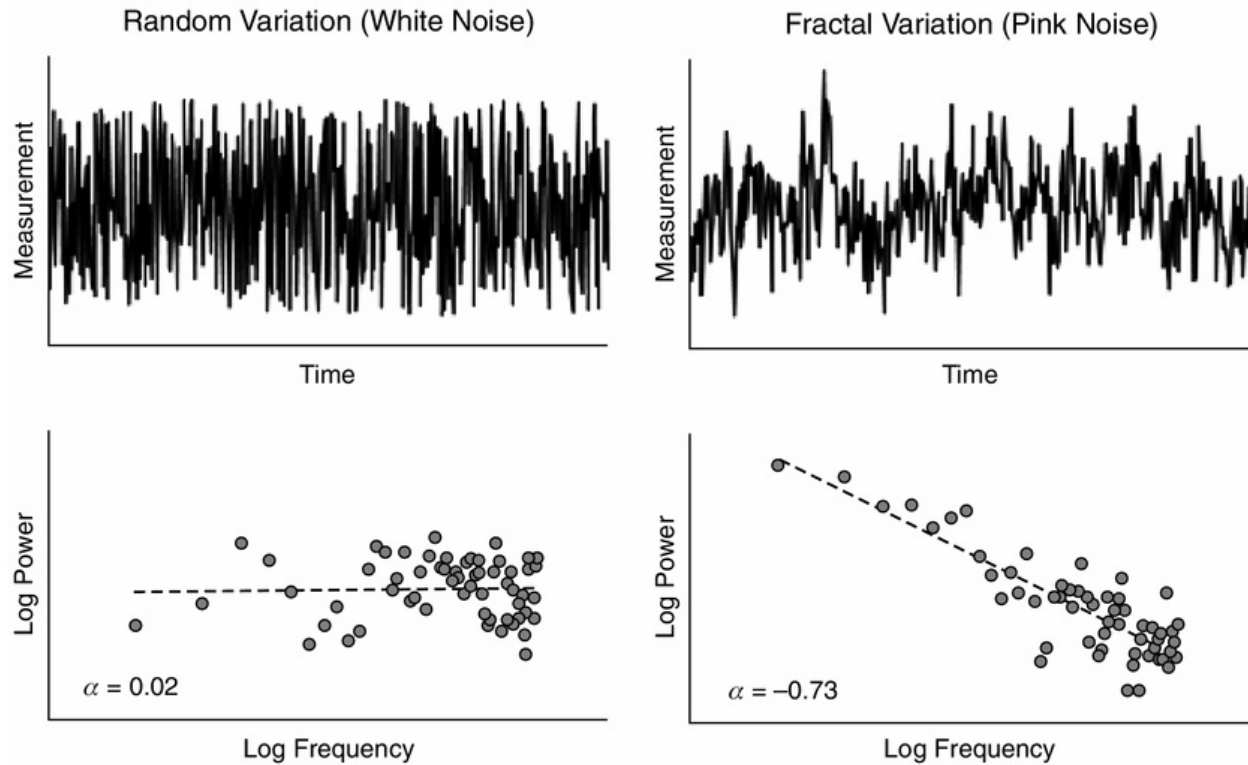


Figure 11.11. Examples of time series composed of random variation (top left) and fractal variation (top right) and the associated spectral plots with logarithmic axes (bottom left and right, respectively).

A fractal time-series pattern is characterized by an inverse proportional relationship between the power (P) and frequency (f) of observed variation in a time series of measurements.⁶ That is, for a fractal time-series, there exists a proportional relationship between the size of a change and how frequently changes of that size occur, with this relationship remaining stable across changes in scale. It is in this sense that the pattern of variability in a repeatedly measured behavior is self-similar; large-scale changes occur with the same *relative* frequency as small-scale changes. The degree to which a dataset approximates this ideal relationship between power and frequency, $P = 1/f^\alpha$, is summarized in the scaling exponent, α , with P = power, and f = frequency. If one plots the power of the different spectral frequencies that make up a time series on double-logarithmic axes, α is equivalent to the slope of the line that best fits the data (see Figure 11.11). That is, α captures the relationship between size and frequency of fluctuations in the time series of behavior. Random fluctuations (i.e., white noise) produce a flat line in a log-log spectral plot with a slope close to 0, which indicates that changes of all different sizes occur with approximately

the same frequency in the time series. Alternatively, fractal fluctuations, often referred to as pink or $1/f$ noise, produce a line in a log-log spectral plot that has a slope closer to -1 , which indicates the self-similar and scale-invariant scaling relationship characteristic of fractal patterns.

It is becoming increasingly clear that the behavior of most natural systems, including human and social systems, exhibit varying degrees of fractal structure (Delignières et al., 2006; Gilden, 2009; Holden, 2005). Moreover, the degree to which the fluctuations in a behavioral time series are fractal (i.e., pink) or not (i.e., white) provides evidence that a system is nonlinear and that its behavior is a consequence of interaction-dominant dynamics (Van Orden et al., 2003). To this extent, fractal patterns of behavior can also be a sign of emergence and self-organization. Although these ideas may seem foreign and irrelevant to the uninitiated, these analyses of fractal fluctuations have been applied fruitfully in the social domain. For example, the behavioral waves or periodic flow of social interaction has a fractal structure (Newton, 1994), as do the dynamics of self-esteem (Delignières et al., 2004). More recently, Correll (2008) has shown that participants who are trying to avoid racial bias show a lesser fractal signature in their response latencies in a video game. Correll discusses these findings in light of characterizing social perception and other processes as a system of many intertwined dependencies – as processes of a complex dynamical system. So the behavioral fluctuations a person gives off may hint at social judgment events, such as stereotyping or racial bias. This avenue of research still seems relatively unexplored, and surprisingly so, if Correll's (2008) results are robust in multiple contexts.

Further Reading

Finally, we should note that there are a number of reviews of this material that can be consulted in social psychology (e.g ., Nowak & Vallacher, 1998; Vallacher & Nowak, 1994) or within cognitive science more broadly (e.g ., Port & Van Gelder, 1995; Spivey, 2007; Warren, 2006). Note as well that there are, of course, different levels of theoretical commitment to this agenda. For example, many dynamical systems approaches explicitly avoid the use of mental representations, and instead focus on perception-action couplings as a basis for understanding social and human behavior (e.g., Marsh et al., 2009; Richardson et al., 2009). Others focus more on the dynamics of internal mental processes or cognitive dynamics (e.g., Spivey, 2007). This leads to theoretical subtleties that cannot be conveyed here. If readers are intrigued to look further, we would

encourage them to decide whether the theory, modeling, or analysis of dynamical systems motivates their interest. If it is deeply theoretical, for example, then much of the work we review would help tease apart those theoretical nuances. We also encourage readers to read the books of [Kelso \(1995\)](#), Nowak and Vallacher (1998), and Thelen and Smith (1994) for a more thorough, but still easy-to-read, introduction to dynamical systems theory and practice. If one is more interested in mathematical modeling, a review of the primary modeling papers cited will help the researcher jump straight into the concepts and their practicality in modeling. Reviews by Kaplan and Glass (1995) and Strogatz (1994) also provide a good introduction to the mathematical details of dynamical systems modeling. With respect to dynamical systems analysis, direct consultation with research that has previously employed a method of interest is always the best place to start. Throughout this chapter we have therefore included a range of relevant research articles that will enable any interested reader to gain a foothold on the relevant literature.

Conclusion

We had three key goals in this chapter. The first was to lay out some of the basic concepts behind complex dynamical systems. These concepts motivate theorists as they seek to understand social and cognitive systems as systems sustained by self-organization, bringing about soft-assembled processes, through nonlinear interaction-dominant dynamics. Our second key goal was to demonstrate how the dynamics of social processes and behavior can be explored directly using mathematical models. By laying out some of the mathematical modeling techniques that can be employed to understand dynamical systems, we showcased how the relatively new concepts of dynamical systems gain concrete manifestation in these explicit models. Our third goal was to provide a brief description of how the dynamics of behavior can be explored via the dynamical analyses of recorded data. Thus, in the third section of this chapter we described just a few of the many linear and nonlinear time-series analyses techniques that a social researcher could employ to investigate the dynamics inherent to his or her own behavioral (time-series) data.

Collectively, these sections urge social-personality psychologists to think of behavior as something that continually *changes* and, therefore, that must be studied and modeled as *time-evolving*. It is often challenging for a researcher to conceptualize his or her context of study in such a way that time series can be collected (see, for example, Correll's 2008 clever use of video games), or to

adapt the perhaps unfamiliar concepts of self-organization or soft-assembly to their theories. Doing so, however, can help us understand the *processes* by which social behaviors come about in day-to-day activities and, thus, the approach will no doubt pay significant dividends for researchers interested in unveiling new domains of inquiry.

Acknowledgments

We thank Charles Coey, Steve Harrison, Alex Demos, Ben Meagher, and, most notably, Rachel W. Kallen for their helpful comments and suggestions. Dr. Richardson was supported, in part, by funding from the National Science Foundation (BCS-092662). Dr. Dale was supported, in part, by NSF BCS-0926670.

References

- Abarbanel, H. D. I. (1996). *Analysis of observed chaotic data*. New York: Springer.
- Abraham, F. D., Abraham, R., & Shaw, C. D. (1990). *A visual introduction to dynamical systems theory for psychology*. Santa Cruz, CA: Aerial Press.
- Anderson, M. L., Richardson, M. J., & Chemero, A. (2012). Eroding the boundaries of cognition: implications of embodiment. *Topics in Cognitive Science*. doi: 10.1111/j.1756-8765.2012.01211. Angus, D., Smith, A., & Wiles, J. (2011). Conceptual recurrence plots: Revealing patterns in human discourse. *IEEE Transactions on Visualization and Computer Graphics*, 18, 988–997.
- Arrow, H. (1997). Stability, bistability, and instability in small group influence patterns. *Journal of Personality and Social Psychology*, 72, 75–85.
- Arrow, H., & McGrath, J. E., & Berdahl, J. L. (2000). *Small groups as complex systems: Formation, coordination, development, and adaptation*. Thousand Oaks, CA: Sage.
- Asch, S. (1952). *Social psychology*. New York: Prentice-Hall.
- Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.
- Axelrod, R., & Dion, D. (1988). The further evolution of cooperation. *Science*, 242, 1385–1390.

- Axelrod, R., & Hamilton, W. (1981). The evolution of cooperation. *Science*, 211, 1390–1396.
- Axelrod, R., Riolo, R. L., & Cohen, M. D. (2002). Beyond geography: Cooperation with persistent links in the absence of clustered. *Personality and Social Psychology Review*, 6, 341–346.
- Bakeman, R., Deckner, D. F., & Quera, V. (2005). Analysis of behavioral streams. In D. M. Teti (Ed.), *Handbook of research methods in developmental science*. Oxford: Blackwell.
- Balcetis, E. E., & Lassiter, G. (2010). *Social psychology of visual perception*. New York: Psychology Press.
- Baron, R. M., Amazeen, P. G., & Beek, P. J. (1994). Local and global dynamics of social relations. In R. R. Vallacher & A. Nowak (Eds.), *Dynamical systems in social psychology* (pp. 111–138). New York: Academic Press
- Barton, S. (1994). Chaos, self-organization, and psychology. *American Psychologist*, 49, 5–14.
- Bassingthwaite, J. B., Liebovitch, L. S., & West, B. J. (1994). *Fractal physiology*. New York: Oxford University Press.
- Bertuglia, C. S., & Vaio, F. (2005). *Nonlinearity, chaos, and complexity: The dynamics of natural and social systems*. Oxford: Oxford University Press
- Boccara, N. (2003) *Modeling complex systems*. New York: Springer.
- Boker, S. M., & Wenger, M. J. (2007). *Data analytic techniques for dynamical systems in the social and behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Brown, K. W., & Moskowitz, D. S. (1998): Dynamic stability: The rhythms of our daily lives. *Journal of Personality*, 66, 105–134.
- Buder, E. H. (1991). A nonlinear dynamic model of social interaction. *Communication Research*, 18, 174–198.
- Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2007). Psychophysiological science: Interdisciplinary approaches to classic questions about the mind. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed., pp. 1–18). New York: Cambridge University Press.

- Camazine, S., Deneubourg, J. L., Franks, N. R., Sneyd, J., Theraulaz, G., & Bonabeau, E. (2001). *Self-organization in biological systems*. Princeton, NJ: Princeton University Press.
- Castellano, G., Kessous, L., & Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech. *Affect and Emotion in Human-Computer Interaction, Lecture Notes in Computer Science*, 4868, 92–103.
- Coleman, P. T., Vallacher, R., Nowak, A., & Bui-Wrzosinska, L. (2007). Intractable conflict as an attractor: Presenting a dynamical model of conflict, escalation, and intractability. *American Behavioral Scientist*, 50, 1454–1475.
- Correll, J. (2008). 1/f noise and effort on implicit measures of racial bias. *Journal of Personality & Social Psychology*, 94, 48–59.
- Cummings, A., Schlosser, A., & Arrow, H. (1996). Developing complex group products: Idea combination in computer-mediated and face-to-face groups. *Computer Supported Cooperative Work*, 4, 229–251.
- Dale, R., & Spivey, M. J. (2005). Categorical recurrence analysis of child language. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, 530–535.
- Dale, R., & Spivey, M. J. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, 56, 391–430.
- Dale, R., Warlaumont, A. S., & Richardson, D. C. (2011). Nominal cross recurrence as a generalized lag sequential analysis for behavioral streams. *International Journal of Bifurcation and Chaos*, 21, 1153–1161.
- Delignières, D., Fortes, M., & Ninot, G. (2004). The fractal dynamics of self-esteem and physical self. *Nonlinear Dynamics in Psychology and Life Science*, 8, 479–510.
- Delignières, D., Ramdani, S., Lemoine, L., Torre, K., Fortes, M., & Ninot, G. (2006). Fractal analysis for short time series: A reassessment of classical methods. *Journal of Mathematical Psychology*, 50, 525–544.
- Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, 130, 355–391.

- Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE*, 6, e26752.
- Dooley, K. J. (1994). A complex adaptive systems model of organization change. *Nonlinear Dynamics, Psychology, and Life Sciences*, 1, 69–97.
- Eguíluz, V. M., Zimmermann, M. G., Cela-Conde, C. J., & San Miguel M. (2005). Cooperation and emergence of role differentiation in the dynamics of social networks. *American Journal of Sociology*, 110, 977–1008.
- Feldman, R., Magori-Cohen, R., Galili, G., Singer, M., & Louzoun, Y. (2011). Mother and infant coordinate heart rhythms through episodes of interaction synchrony. *Infant Behavioral Development*, 34, 569–577.
- Fowler, C. A., Richardson, M. J., Marsh, K. L., & Shockley, K. D. (2008). Language use, coordination, and the emergence of cooperative action. In A. Fuchs & V. Jirsa (Eds.), *Coordination: Neural, behavioral and social dynamics*. (pp. 261–280). Heidelberg: Springer-Verlag.
- Frank, T. D., & Richardson, M. J. (2010). On a test statistic for the Kuramoto order parameter of synchronization: With an illustration for group synchronization during rocking chairs. *Physica D*, 239, 2084–2092.
- Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, 42, 226–241.
- Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, 118, 247–279.
- Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, 2, article 59.
- Gallagher, R., & Appenzeller, T. (1999). Beyond reductionism. *Science*, 284, 79.
- Gilden, D. L. (2001). Cognitive emissions of 1/f noise. *Psychological Review*, 108, 33–56.
- Gilden, D. L. (2009). Global model analysis of cognitive variability. *Cognitive Science*, 33, 1441–1467.
- Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *IEEE International*

- Conference on Acoustics, Speech, and Signal Processing*, 1, 517–520.
- Goldstone, R. L., & Gureckis, T. M. (2009). Collective behavior. *Topics in Cognitive Science*, 1, 412–438.
- Gordon, D. M. (2007). Control without hierarchy. *Nature*, 446, 143.
- Gottman, J., Murray, J., Swanson, C., Tyson, R., & Swanson, K., (2003). *The mathematics of marriage: Dynamic nonlinear models*. Cambridge, MA: MIT Press.
- Gottman, J., Swanson, C., & Swanson, K., (2002). A general systems theory of marriage: Nonlinear difference equation modeling of marital interaction. *Personality and Social Psychology Review*, 6, 326–340.
- Gottman, J. M. (1981). *Time series analysis: A comprehensive introduction for social scientist*. Cambridge: Cambridge University Press.
- Gottschalk, A., Bauer, M. S., & Whybrow, P. C. (1995). Evidence of chaotic mood variation in bipolar disorder. *Archives of General Psychiatry*, 52, 947–959.
- Guastello, S. J. (1995). *Chaos, catastrophe, and human affairs: Applications of nonlinear dynamics to work, organizations, and social evolution*. Mahwah, NJ: Lawrence Erlbaum.
- Guastello, S. J. (2002). *Managing emergent phenomena: Nonlinear dynamics in work organizations*. Mahwah, NJ: Erlbaum.
- Guastello, S. J., Koopmans, M., & Pincus, D. (2009). *Chaos and complexity in psychology: Theory of nonlinear dynamics*. New York: Cambridge University Press.
- Guastello, S. J., & Liebovitch, L. S. (2009). Introduction to nonlinear dynamics and complexity. In S. J. Guastello, M. Koopmans, & D. Pincus (Eds.), *Chaos and complexity in psychology: Theory of nonlinear dynamics* (pp. 1–40). New York: Cambridge University Press
- Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, 51, 347–356.
- Harrison, S. J., & Richardson, M. J. (2009). Horsing around: Spontaneous four-legged coordination. *Journal of Motor Behavior*, 41, 519–524.

- Heath, R. A. (2000). *Nonlinear dynamics: Techniques and applications in psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Holden, J. G. (2005). Gauging the fractal dimension of response times from cognitive tasks. In M. A. Riley & G. C. Van Orden (Eds.), *Contemporary nonlinear methods for behavioral scientists: A webbook tutorial* (pp. 267–318). Retrieved April 8, 2005, from <http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp>
- Iberall, A., & McCulloch, W. (1969). The organizing principle of complex living systems. *Journal of Basic Engineering*, 91, 290–294.
- Isenhower, R., Richardson, M. J., Marsh, K. L., Carello, C., & Baron, R. M. (2010). Affording cooperation: Dynamics and action-scaled invariance of joint lifting. *Psychonomic Bulletin and Review*, 17, 342–347.
- Kantz, H., & Schreiber, T. (1997/2003) *Nonlinear time series analysis*. Cambridge: Cambridge University Press.
- Kaplan, D., & Glass, L. (1995). *Understanding nonlinear dynamics*. New York: Springer-Verlag.
- Kello, C. T., Beltz, B. C., Holden, J. H., & Van Orden, G. C. (2007). The emergent coordination of cognitive function. *Journal of Experimental Psychology: General*, 136, 551–568.
- Kelso, J. A. S. (1995). *Dynamic patterns*. Cambridge, MA: MIT Press.
- Kelso, J. A. S. (2009). Synergies: Atoms of brain and behavior. In D. Sternad (Ed.), *Progress in motor control* (pp. 83–91). Heidelberg, Germany: Springer.
- Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford Press.
- Kreuz, R. J., & Riordan, M. A. (2011). The transcription of face-to-face interaction. In W. Bublitz & N. R. Norrick (Eds.), *Foundations of pragmatics* (pp. 657–680). New York: De Gruyter Mouton.
- Larsen, R. J., & Kasimatis, M. (1990). Individual differences in entrainment of mood to the weekly calendar. *Journal of Personality and Social Psychology*, 58, 164–171.
- Lewin, K. (1936). *Principles of topological psychology*. New York: McGraw-Hill.

- Lewin, K., Lippitt, R., & White, R. K. (1939). Patterns of aggressive behavior in experimentally created “social climates.” *Journal of Social Psychology*, 10, 271–299.
- Loehr, D., & Harper, L. (2003). Commonplace tools for studying commonplace interactions: Practitioners’ notes on entry-level video analysis. *Visual Communication*, 2, 225–233.
- Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, 36, 1404–1426.
- MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Mandelbrot, B. B. (1982). *The fractal geometry of nature*. San Francisco: W. H. Freeman & Co.
- Marsh, K. L. (2010). Sociality from an ecological, dynamical perspective. In G. R. Semin & G. Echterhoff (Eds.), *Grounding sociality: Neurons, minds, and culture* (pp. 43–71). London: Psychology Press.
- Marsh, K. L., Richardson, M. J., & Schmidt, R. C. (2009). Social connection through joint action and interpersonal coordination. *Topics in Cognitive Science*, 1, 320–339.
- Marwan, N. (2008). A historical review of recurrence plots. *European Physical Journal*, 164, 3–12.
- Marwan, N., & Meinke, A. (2002). Extended recurrence plot analysis and its application to ERP data. *International Journal of Bifurcation and Chaos*, 14, 761–771.
- Mason, W. A., Conrey, F. R., & Smith, E. R. (2007). Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and Social Psychology Review*, 11, 279–300.
- Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, 5, 175–190.
- McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing*, Vol. 2. Cambridge, MA: MIT Press.

- McGarva, A., & Warner, R. M. (2003). Attraction and social coordination: Mutual entrainment of vocal activity rhythms. *Journal of Psycholinguistic Research*, 32, 335–354.
- Mead, G. H. (1934). *On social psychology*. Chicago: University of Chicago Press.
- Meadows, D. H. (2008). *Thinking in systems: A primer*. White River Junction, VT: Chelsea Green Publishing.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806–834.
- Miles, L. K., Lumsden, J., Richardson, M. J., & Macrae, N. C. (2011). Do birds of a feather move together? Group membership and behavioral synchrony. *Experimental Brain Research*, 3–4, 495–503.
- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102, 246–268.
- Mitra, S., & Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 37, 311–324.
- Moussaïd, M., Perozo, N., Garnier, S., Helbing, D., & Theraulaz, G. (2010). The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PLoS ONE*, 5, e10047.
- Newtson, D. (1994) The perception and coupling of behavior waves. In R. Vallacher & A. Nowak (Eds.), *Dynamical systems in social psychology* (pp. 139–167). San Diego, CA: Academic Press.
- Nowak, A., & Lewenstein, M. (1996) Modelling social change with cellular automata. In R. Hegselmann, K. Troitzsch, & U. Muller (Eds.), *Computer simulations from the philosophy of science point of view*. Dordrecht: Kluwer.
- Nowak, A., & Vallacher, R. R. (1998). *Dynamical social psychology*. New York: Guilford Press.
- Nowak, A., Vallacher, R. R., & Borkowski, W. (2000). Modeling the temporal coordination of behavior and internal states. In G. Ballot & G. Weisbuch

- (Eds.), *Application of simulations to social sciences* (pp. 67–86). Oxford: Hermes Science Publications.
- Porges, S. W., Bohrer, R. E., Cheung, M. N., Drasgow, F., McCabe, P. M., & Keren, G. (1980). New time-series statistic for detecting rhythmic co-occurrence in the frequency domain: The weighted coherence and its application to psychophysiological. *Psychological Bulletin*, 88, 580–587.
- Port, R. F., & Van Gelder, T. (1995). *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.
- Read, S. J., & Miller, L. C. 2002. Virtual personalities: A neural network model of personality. *Personality and Social Psychology Review*, 6, 357–369.
- Reis, H. T. (2008). Reinvigorating the concept of situation in social psychology. *Personality and Social Psychology Review*, 12, 311–329.
- Richardson, D. C & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, 29, 1045–1060.
- Richardson, D. C., Dale, R., & Spivey, M. J. (2007). Eye movements in language and cognition. In M. Gonzalez-Marquez, I. Mittelberg, S. Coulson, & M. J. Spivey (Eds.), *Methods in cognitive linguistics* (pp. 328–341). Amsterdam/Philadelphia: John Benjamins.
- Richardson, D. C., Dale, R., & Tomlinson, J. M. (2009). Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science*, 33, 1468–1482.
- Richardson, M. J., Fajen, B. R., Shockley, K., Riley, M. A., & Turvey, M. T. (2008). Ecological psychology: Six principles for an embodied-embedded approach to behavior. In P. Calvo & T. Gomila (Eds.), *Handbook of cognitive science: An embodied approach* (pp. 161–197). San Diego, CA: Elsevier.
- Richardson, M. J., Garcia, A., Frank, T. D., Gergor, M., & Marsh, K. L. (2012). Measuring group synchrony: A cluster-phase method for analyzing multivariate movement time-series. *Frontiers in Physiology*, 3, 405.
- Richardson, M. J., Marsh, K. L., Isenhower, R., Goodman, J., & Schmidt, R. C. (2007). Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human Movement Science*, 26, 867–891.

- Richardson, M. J. Marsh, K. L., & Schmidt, R. C. (2010). Challenging egocentric notions of perceiving, acting, and knowing. In L. F. Barrett, B. Mesquita, & E. Smith (Eds.), *The mind in context* (pp. 307–333). New York: Guilford.
- Richardson, M. J., Schmidt, R. C., & Kay, B. A. (2007). Distinguishing the noise and attractor strength of coordinated limb movements using recurrence analysis. *Biological Cybernetics*, 96, 59–78.
- Riley, M. A., Balasubramaniam, R., & Turvey, M. T. (1999) Recurrence quantification analysis of postural fluctuations. *Gait & Posture*, 9, 65–78.
- Riley, M. A., & Van Orden, G. C. 2005. *Tutorials in contemporary nonlinear methods for the behavioral sciences*. Retrieved April 8, 2005, from <http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp>.
- Rinaldi, S., & Gragnani, A. (1998). Love dynamics between secure individuals: A modeling approach. *Nonlinear Dynamics, Psychology, and Life Sciences*, 2, 283–301.
- Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing, Vol. 1*. Cambridge, MA: MIT Press.
- Sadler, P., Ethier, N., Gunn, G. R., Duong, D., & Woody, E. (2009). Are we on the same wavelength? Complementarity as shared cyclical patterns within an interaction. *Journal of Personality & Social Psychology*, 97(6), 1005–1020.
- Sadler, P., Ethier, N., & Woody, E. (2011). Interpersonal complementarity. In L. M. Horowitz & S. N. Strack (Eds.), *Handbook of interpersonal psychology: Theory, research, assessment, and therapeutic interventions* (pp. 123–142). New York: Wiley.
- Schmidt, R. C., Carello, C., & Turvey, M. T. (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 227–247.
- Schmidt, R. C., Fitzpatrick, P., Caron, R., & Mergeche, J. (2011). Understanding social motor coordination. *Human Movement Science*, 30, 834–845.
- Schmidt, R. C., Morr, S., Fitzpatrick, P., & Richardson, M. J. (2012). Measuring the dynamics of interactional synchrony. *Journal of Nonverbal Behavior*. 36, 263–279.

- Schmidt, R. C., & O'Brien, B. (1997). Evaluating the dynamics of unintended interpersonal coordination. *Ecological Psychology*, 9, 189–206.
- Schmidt, R. C., & Richardson, M. J. (2008). Dynamics of interpersonal coordination. In A. Fuchs & V. Jirsa (Eds.), *Coordination: Neural, behavioural and social dynamics* (pp. 281–308). Heidelberg: Springer-Verlag.
- Schuldberg, D., & Gottlieb, J. (2002). Dynamics and correlates of microscopic changes in affect. *Nonlinear Dynamics, Psychology and Life Sciences*, 6, 231–257.
- Semin, G. R., & Smith, E. R. (2008). *Embodied grounding: Social, cognitive, affective, and neuroscientific approaches*. New York: Cambridge University Press.
- Shelling, T. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1, 143–186.
- Shockley, K., Santana, M. V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, 29, 326–332.
- Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of Personality and Social Psychology*, 70, 893–912.
- Smith, E. R., & Collins, E. C. (2009). Contextualizing person perception: Distributed social cognition. *Psychological Review*, 116, 343–364.
- Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, 11, 87–104.
- Smith, E. R., & DeCoster, J. (1998). Person perception and stereotyping in a recurrent connectionist network using distributed representations. In S. J. Read & L. Miller (Eds.), *Connectionist and parallel distributed processing models of social reasoning and behavior* (pp. 111–140). Hillsdale, NJ: Lawrence Erlbaum.
- Spivey, M. J. (2007). *The continuity of mind*. Oxford: Oxford University Press.
- Sprott, J. C. (2004). Dynamical models of love. *Nonlinear Dynamics in Psychology and the Life Sciences*, 8, 303–314.
- Stam, C. J. (2005). Nonlinear dynamical analysis of EEG and MEG: Review of

- an emerging field. *Clinical Neurophysiology*, 116, 2266–2301.
- Stephen, D. G., & Mirman, D. (2010). Interactions dominate the dynamics of visual cognition. *Cognition*, 115, 154–165.
- Strogatz, S. H. (1994). *Nonlinear dynamics and chaos: With applications to physics, chemistry and engineering*. Cambridge, MA: Addison-Wesley.
- Sumpter, D. J. T. (2010) *Collective animal behavior*. Princeton, NJ: Princeton University Press.
- Tesser, A. (1980). When individual dispositions and social pressure conflict: A catastrophe. *Human Relations*, 33, 393–407.
- Tesser, A., & Achee, J. (1994). Aggression, love, conformity, and other social psychological catastrophes. In R. R. Vallacher & A. Nowak (Eds.), *Dynamical systems in social psychology* (pp. 96–109). San Diego, CA: Academic.
- Theiner, G., Allen, C., Goldstone, R. L., (2010). Recognizing group cognition. *Cognitive Systems Research*, 11, 378–395.
- Thelen, E., & Smith, L. B. (1994). *A dynamical systems approach to the development of cognition and action*. Cambridge, MA: Bradford-MIT Press.
- Theraulaz, G., & Bonabeau, E. (1995). Coordination in distributed building. *Science*, 269, 686–688.
- Tognoli, E., Lagarde, J., DeGuzman, C., & Kelso, J. A. S. (2007). The phi complex as a neuromarker of human social coordination. *PNAS*, 104, 8190–8195.
- Tuller, B. (2004). Categorization and learning in speech perception as dynamical processes. In M. A. Riley & G. C. Van Orden (Eds.), *Tutorials in contemporary nonlinear methods for the behavioral sciences*. Retrieved from <http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp>
- Tuller, B., Case, P., Ding, M., & Kelso, J. A. S. (1994). The nonlinear dynamics of speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, 20, 1–14.
- Vallacher, R. R., & Nowak, A. (1994). *Dynamical systems in social psychology*. San Diego, CA: Academic Press.
- Vallacher, R. R., Nowak, A., & Zochowski, R. (2005). Dynamics of social

coordination: The synchronization of internal states in close relationships.
Retrieved 2009 from
[http://psy2.fau.edu/~vallacher/pdfs/articles/Vallacher_et_al_\(2005\)_Dynamics](http://psy2.fau.edu/~vallacher/pdfs/articles/Vallacher_et_al_(2005)_Dynamics)

Vallacher, R. R., Read, S. J., & Nowak, A. (2002). The dynamical perspective in personality and social psychology. *Personality and Social Psychology Review*, 6, 264–273.

van der Mass, H. L. J., Kolstein, R., & van der Pligt, J. (2003). Sudden transitions in attitudes. *Sociological Methods & Research*, 32, 125–152.

van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review*, 98, 3–53.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, 132, 331–350.

Van Orden, G. C., Kello, C. T., & Holden, J. G. (2010). Situated behavior and the place of measurement in psychological theory. *Ecological Psychology*, 22, 24–43.

Van Orden, G. C., Kloos, H., & Wallot, S. (2011). Living in the pink: Intentionality, wellbeing, and complexity. In C. A. Hooker (Ed.), *Philosophy of complex systems: Handbook of the philosophy of science* (pp. 639–684). Amsterdam: Elsevier.

Van Orden, G. C., & Stephen, D. G. (2012). Cognitive science usefully cast as complexity science? *Topics in Cognitive Science*, 4, 3–6.

von Holst, E. (1939/1973). Relative coordination as a phenomenon and as a method of analysis of central nervous system function. In R. Martin (Ed. and Trans.), *The collected papers of Erich von Holst: Vol. 1. The behavioral physiology of animal and man* (pp. 33–135). Coral Gables, FL: University of Miami Press.

Warlaumont, A. S., Oller, D. K., Dale, R., Richards, J. A., Gilkerson, J., & Dongxin, X. (2010). Vocal interaction dynamics of children with and without autism. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 121–126). Austin, TX: Cognitive Science Society.

Warner, R. M. (1992). Sequential analysis of social interaction: Assessing

internal versus social determinants of behavior. *Journal of Personality and Social Psychology*, 63, 51–60.

Warner, R. M. (1998). *Spectral analysis of time-series data*. New York: Guilford Press.

Warner, R. M., Waggener, T. B., & Kronauer, R. E. (1983). Synchronized cycles in ventilation and vocal activity during spontaneous conversational speech. *Journal of Applied Physiology: Respiratory, Environmental and Exercise Physiology*, 54, 1324–1334.

Warren, W. H. (2006). The dynamics of perception and action. *Psychological Review*, 113, 358–389.

Webber, Jr., C. L., & Zbilut, J. P. (2005). Recurrence quantification analysis of nonlinear dynamical systems. In M. A. Riley & G. C. Van Orden (Eds.), *Contemporary nonlinear methods for behavioral scientists: A webbook tutorial* (pp. 26–94). Retrieved April 8, 2005, from <http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp>

Wolfram, S. (2002). *A new kind of science*. Champaign, IL: Wolfram Media.

Wyer, R. S., & Srull, T. K. (1989). *Memory and cognition in its social context*. Hillsdale, NJ: Erlbaum.

Zbilut, J. P., Thomasson, N., & Webber, C. L. (2002). Recurrence quantification analysis as a tool for nonlinear exploration of nonstationary cardiac signals. *Medical Engineering & Physics*, 24, 53–60.

¹ It is a nonlinear equation because if we expand $rx_{(t)}(1 - x_{(t)})$, we get $r(x_{(t)} - x_{(t)}^2)$, such that the state of the system is the product of a constant and the state variable to a power greater than one (e.g., x^2 , x^3 , $x^4 \dots$), in this case $x_{(t)}^2$.

² A cellular automaton could also be a three-dimensional cube of cells, for instance.

³ The converse is also true, with some researchers focusing primarily on building abstract dynamical models like those described in the previous section

without collecting real behavioral data (time-series or otherwise). Just as valid as empirical research, this latter “research by modeling” approach is less concerned with simulating real-world behavior and is more concerned with developing a formal yet highly generalizable understanding of how organized behavior can emerge, change, and dissolve over time.

⁴ Subtle variations in time interval are not always catastrophic, especially for longer intervals (i.e., hour or day), but should be minimized as much as possible.

⁵ In actuality, only ideal mathematical or geometric fractals are truly self-similar, with real-world fractals considered self-similar in a statistical sense. *Statistical self-similarity* simply means that a pattern is composed of statistically similar copies of itself (looks, on average, similar at different scales of observation).

⁶ For an introductory review of the various methods that can be employed to measure the fractal structure of time-series data, see Delignières *et al.* (2006).

Chapter twelve Implicit Measures in Social and Personality Psychology

Bertram Gawronski and Jan De Houwer

Self-report measures arguably represent one of the most important research tools in social and personality psychology. To measure people's attitudes, beliefs, and personality characteristics, it seems rather straightforward to simply ask them about their thoughts, feelings, and behaviors. Yet, researchers are well aware that people are sometimes unwilling or unable to provide accurate reports of their own psychological attributes. In socially sensitive domains, for example, responses on self-report measures are often distorted by social desirability and self-presentational concerns. Similarly, the value of self-report measures seems limited for psychological attributes that are introspectively inaccessible or outside of conscious awareness. To overcome these limitations, psychologists have developed alternative measurement instruments that reduce participants' ability to control their responses and do not require introspection for the assessment of psychological attributes. In social and personality psychology, such measurement instruments are commonly referred to as *implicit measures*, whereas traditional self-report measures are often described as *explicit measures*.

The main goal of the current chapter is to provide a general introduction to the use and meaning of implicit measures in social and personality psychology. Toward this end, we first explain what implicit measures are and in which sense they may be described as implicit. We then provide an overview of the currently available paradigms, including descriptions of their basic procedures and some recommendations on how to choose among the various measures. Expanding on this overview, we outline the kinds of insights that can be gained from implicit measures for understanding the determinants of behavior, biases in information processing, and the formation and change of mental representations. In the final two sections we discuss some caveats regarding the interpretation of implicit measures and potential directions for future developments.

What Are Implicit Measures?

A central characteristic of implicit measures is that they aim to capture psychological attributes (e.g., attitudes, stereotypes, self-esteem) without requiring participants to report a subjective assessment of these attributes. However, there are a lot of such indirect measurement techniques and only a few of them have been described as implicit. Thus, a frequent question in research using implicit measures concerns the meaning of the terms “implicit” and “explicit.” This issue is a common source of confusion, because some researchers use the terms to describe features of measurement procedures, whereas others use them to describe the nature of the psychological attributes assessed by particular measurement instruments. For example, it is sometimes argued that participants are aware of what is being assessed by an explicit measure but they are unaware of what is being assessed by an implicit measure (e.g., Petty, Fazio, & Briñol, 2009). Yet, other researchers assume that the two kinds of measures tap into distinct memory representations, such that explicit measures tap into conscious representations whereas implicit measures tap into unconscious representations (e.g., Greenwald & Banaji, 1995).

Although these conceptualizations are relatively common in the literature on implicit measures, we believe that it is conceptually more appropriate to classify different measures in terms of whether the to-be-measured psychological attribute influences participants’ responses on the task in an automatic fashion (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009). Specifically, measurement outcomes may be described as implicit if the impact of the to-be-measured psychological attribute on participants’ responses is unintentional, resource-independent, unconscious, or uncontrollable. Conversely, measurement outcomes may be described as explicit if the impact of the to-be-measured psychological attribute on participants’ responses is intentional, resource-dependent, conscious, or controllable (cf. Bargh, 1994; Moors & De Houwer, 2006). For example, a measure of racial attitudes may be described as implicit if it reflects participants’ racial attitudes even when they do not have the goal to express these attitudes (i.e., unintentional) or despite the goal to conceal these attitudes (i.e., uncontrollable).

An important aspect of this conceptualization is that the terms “implicit” and “explicit” describe the process by which a psychological attribute influences measurement outcomes rather than the measurement procedure itself (e.g., Petty et al., 2009) or the underlying psychological attribute (e.g., Greenwald & Banaji, 1995). Moreover, whereas the classification of measurement outcomes as implicit or explicit depends on the processes that underlie a given measurement

procedure, measurement procedures may be classified as direct or indirect on the basis of their objective structural properties (De Houwer & Moors, 2010). Specifically, a measurement procedure can be described as direct when the measurement outcome is based on participants' self-assessment of the to-be-measured attribute (e.g., when participants' racial attitudes are inferred from their self-reported liking of black people). Conversely, a measurement procedure can be described as indirect when the measurement outcome is not based on a self-assessment (e.g., when participants' racial attitudes are inferred from their reaction time performance in a speeded categorization task) or when it is based on a self-assessment of attributes other than the to-be-measured attribute (e.g., when participants' racial attitudes are inferred from their self-reported liking of a neutral object that is quickly presented after a black face). In line with this conceptualization, we use the terms "direct" and "indirect" to describe measurement procedures and the terms "explicit" and "implicit" to describe measurement outcomes. However, because claims about the automatic versus controlled nature of measurement outcomes have to be verified through empirical data, descriptions of measures as implicit should be interpreted as tentative (for a review of relevant evidence, see De Houwer et al., 2009). We return to this issue when we discuss caveats regarding the interpretation of implicit measures, in particular the joint contribution of automatic and controlled processes.

An Overview of Basic Paradigms

The use of implicit measures in social and personality psychology has its roots in the mid-1980s, when researchers adopted sequential priming tasks from cognitive psychology to study the automatic activation of attitudes (Fazio, Sanbonmatsu, Powell, & Kardes, 1986) and stereotypes (Gaertner & McLaughlin, 1983). These studies provided the foundation for the development of Greenwald, McGhee, and Schwartz's (1998) implicit association test (IAT), which stimulated the current surge in the use of implicit measures. Over the past decade, the toolbox of available measurement instruments has grown substantially through the development of new paradigms and the refinement of existing tasks. In the following sections we provide an overview of the currently available paradigms, including details on their task structure, reliability, and applicability.¹

Implicit Association Test

One of the most frequently used paradigms is Greenwald et al.'s (1998) IAT. The IAT consists of two binary categorization tasks that are combined in a manner that is either compatible or incompatible with the to-be-measured psychological attributes. For example, in an IAT to assess preferences for white over black people, participants are successively presented with positive and negative words and pictures of black and white faces that have to be classified as positive and negative or as black and white, respectively. In one of the two critical blocks, the two categorization tasks are combined in such a way that participants have to respond to positive words and pictures of white faces with one key and to negative words and pictures of black faces with another key. In the other critical block, participants have to respond to positive words and pictures of black faces with one key and to negative words and pictures of white faces with another key. The basic idea underlying the IAT is that quick and accurate responses are facilitated when the key mapping in the task is compatible with a participant's preference (e.g., black-negative; white-positive) but impaired when the key mapping is preference-incompatible (e.g., white-negative; black-positive). Based on this consideration, the mean difference in participants' response latency (or error rates) in the two blocks is typically interpreted as an index of their preference for white over black people, or vice versa depending on the calculation of the difference score (for details regarding the scoring of IAT data, see Greenwald, Nosek, & Banaji, 2003).

A typical IAT includes a total of five blocks. Two of the five blocks contribute the critical trials for the calculation of the so-called IAT score; the other three blocks include practice trials for the two critical blocks (see Table 12.1). For example, an IAT to measure preferences for white over black people would begin with a first practice block in which participants are asked to categorize pictures of black and white faces as fast and accurately as possible as black versus white (*initial target-concept discrimination*). In a second practice block, participants are presented with positive and negative words that must be categorized as pleasant versus unpleasant, again as quickly and accurately as possible (*initial attribute discrimination*). In the third block, the two categorization tasks are combined, such that participants are presented with words and pictures in alternating order, which must be categorized according to the same key assignments as in the first two blocks (*initial combined task*). For example, participants may be asked to press a right-hand key every time they see a positive word or a picture of a white person and a left-hand key every time they see a negative word or a picture of a black person. As with the first two blocks, participants are asked to respond as quickly and accurately as possible.

The fourth block is almost equivalent to the first block, the only difference being that the key assignment for the two target categories is now reversed (*reversed target-concept discrimination*). Finally, the fifth block again combines the two categorization tasks, this time using the key assignments of the second and fourth blocks (*reversed combined task*). In the current example, this would imply that participants have to press a right-hand key every time they see a positive word or a picture of a black person and a left-hand key every time they see a negative word or a picture of a white person.

Table 12.1. Task Structure of an Implicit Association Test (Greenwald et al., 1998) Designed to Assess Preferences for Whites over Blacks (Race-IAT)

Block	Key Assignment			
	Compatible-Incompatible Block Order		Incompatible-Compatible Block Order	
	Left Key	Right Key	Left Key	Right Key
1	Black	White	White	Black
2	Negative	Positive	Negative	Positive
3	Negative/Black	Positive/White	Negative/White	Positive/Black
4	White	Black	Black	White
5	Negative/White	Positive/Black	Negative/Black	Positive/White

The IAT is a very flexible task that can be used to assess almost any type of association between pairs of concepts. For example, by using evaluative attribute dimensions (e.g., pleasant vs. unpleasant), the IAT can be used to assess relative preferences between pairs of objects or categories. Alternatively, the evaluative attribute dimension may be replaced with a semantic dimension to assess semantic associations (e.g., stereotypical associations between black and white people and the attributes of being athletic vs. intelligent). The same flexibility applies to the use of target categories, which may include any pair of objects or categories that can be contrasted in a meaningful manner (e.g., male vs. female). Examples of previous applications include IATs designed to assess prejudice, stereotypes, attitudes toward consumer products, the self-concept, and self-esteem (for an overview, see Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005). Another advantage of the IAT is that it typically shows reliability estimates that are comparable to the ones of traditional self-report measures (see [Table 12.2](#)).²

Nevertheless, the IAT has also been the target of methodological criticism (for

a detailed discussion, see Teige-Mocigemba, Klauer, & Sherman, 2010). A very common concern is that the task structure of the IAT is inherently comparative, which undermines its suitability for the assessment of associations to a single target concept or a single attribute. For example, the race IAT can be used to assess relative preferences for whites over blacks (or vice versa), but it is not possible to calculate separate indices for evaluations of blacks and evaluations of whites (Nosek, Greenwald, & Banaji, 2005). Another concern is that the presentation of compatible and incompatible trials in separate, consecutive blocks can distort measurement scores through various sources of systematic error variance (Teige-Mocigemba et al., 2010). To overcome these shortcomings, researchers have developed a number of procedural variants of the IAT. These variants include modifications that make the IAT amenable for assessing associations of a single target concept (Single Category IAT; Karpinski & Steinman, 2006) or a single attribute (Single Attribute IAT; Penke, Eichstaedt, & Asendorpf, 2006), variants that avoid blocked presentations of compatible and incompatible trials by combining them in a single block (Recoding Free IAT; Rothermund, Teige-Mocigemba, Gast, & Wentura, 2009; Single Block IAT; Teige-Mocigemba, Klauer, & Rothermund, 2008), and an abbreviated variant that is considerably shorter than the standard IAT (Brief IAT; Sriram & Greenwald, 2009). Although the suggested modifications seem quite promising, the currently available evidence is still too scarce to judge whether they retain the functional properties of the standard IAT. The only exception in this regard is the Single Category IAT (Karpinski & Steinman, 2006), which has demonstrated its usefulness in a considerable number of studies.

Evaluative Priming Task

The evaluative priming task employs the basic procedure of sequential priming to assess evaluative responses (Fazio et al., 1986). Toward this end, participants are briefly presented with a prime stimulus (e.g., a black face) that is followed by a positive or negative target word. In the typical version of the task, participants are asked to quickly determine whether the target word is positive or negative by pressing one of two response keys (*evaluative decision task*). To the extent that the prime stimulus leads to faster responses to positive words (compared to a neutral baseline prime), the prime stimulus is assumed to be associated with positive valence. However, if the prime stimulus facilitates responses to negative words (compared to a neutral baseline prime), it is assumed to be associated with negative valence (for details regarding the

calculation of priming scores, see Wittenbrink, 2007). The evaluative priming task can be used to assess evaluative responses to any type of object that can be presented as a prime stimulus in a sequential priming task, and it has been successfully used with prime presentations above the threshold of conscious awareness (i.e., supraliminal presentation) as well as extremely short prime presentations that remain below conscious awareness (i.e., subliminal presentation). Although the standard variant of the task employs evaluative decisions about positive and negative target words, procedural modifications that have been proposed include the pronunciation of positive and negative target words (Bargh, Chaiken, Raymond, & Hymes, 1996) and the naming of positive and negative pictures as target stimuli (Spruyt, Hermans, De Houwer, Vandekerckhove, & Eelen, 2007).

A major advantage of the evaluative priming task is that it allows researchers to calculate separate priming scores for different kinds of associations that are confounded in the IAT (Wittenbrink, 2007). For example, in an evaluative priming task using black and white faces as primes and positive and negative words as targets, the inclusion of a neutral baseline prime (e.g., a grey square) makes it possible to separately measure (a) positive associations with white faces (defined as the difference in response latencies to positive words following white vs. neutral primes), (b) positive associations with black faces (defined as the difference in response latencies to positive words following black vs. neutral primes), (c) negative associations with white faces (defined as the difference in response latencies to negative words following white vs. neutral primes), and (d) negative associations with black faces (defined as the difference in response latencies to negative words following black vs. neutral primes). Although research using the evaluative priming task has provided important insights into the mechanisms underlying attitude-behavior relations (for a review, see Fazio, 2007), a major problem is its low reliability, which rarely exceeds Cronbach's Alpha values of .50 (see Table 12.2).

Table 12.2. Overview of Measurement Procedures, Flexibility of Applications, and Approximate Range of Reliability Estimates

Task	Reference	Applications	Targets	Attributes	Reliability
Action Interference Paradigm	Banse et al. (2010)	(content-specific) ^a	pairs	pairs	.30–.50
Affect Misattribution Procedure	Payne et al. (2005)	evaluative, semantic	individual	pairs	.70–.90
Approach-Avoidance Task	Chen & Bargh (1999)	evaluative	individual	individual	.00–.90 ^b
Brief Implicit Association Test	Sriram & Greenwald (2009)	evaluative, semantic	pairs	pairs	.55–.95
Evaluative Movement Assessment	Brendl et al. (2005)	evaluative	individual	individual	.30–.80 ^c
Evaluative Priming Task	Fazio et al. (1986)	evaluative	individual	individual	.00–.55
Extrinsic Affective Simon Task	De Houwer (2003)	evaluative, semantic	individual	individual	.15–.65
Go/No-go Association Task	Nosek & Banaji (2001)	evaluative, semantic	individual	pairs	.45–.75
Identification Extrinsic Affective Simon Task	De Houwer & De Bruycker (2007)	evaluative, semantic	individual	pairs	.60–.70
Implicit Association Procedure	Schnabel et al. (2006)	self-related	individual	pairs	.75–.85
Implicit Association Test	Greenwald et al. (1998)	evaluative, semantic	pairs	pairs	.70–.90 ^d
Implicit Relational Assessment Procedure	Barnes-Holmes et al. (2010)	evaluative, semantic	individual	individual	.20–.80
Recoding Free Implicit Association Test	Rothermund et al. (2009)	evaluative, semantic	pairs	pairs	.55–.65
Semantic Priming (Lexical Decision Task)	Wittenbrink et al. (1997)	semantic	individual	individual	n/a
Semantic Priming (Semantic Decision Task)	Banaji & Hardin (1996)	semantic	individual	individual	n/a
Single Attribute Implicit Association Test	Penke et al. (2006)	evaluative, semantic	pairs	individual	.70–.80
Single Block Implicit Association Test	Teige-Mocigemba et al. (2008)	evaluative, semantic	pairs	pairs	.60–.90
Single Category Implicit Association Test	Karpinski & Hilton (2006)	evaluative, semantic	individual	pairs	.70–.90
Sorting Paired Features Task	Bar-Anan et al. (2009)	evaluative, semantic	individual	individual	.40–.70

^a Previous applications are limited to gender-stereotyping, although alternative applications seem possible.

^b Reliability estimates differ depending on whether approach-avoidance responses involve valence-relevant or valence-irrelevant categorizations, with valence-irrelevant categorizations showing lower reliability estimates (.00–.35) compared to valence-relevant categorizations (.70–.90).

^c Reliability estimates differ depending on whether the scores involve within-participant comparisons of preferences for different objects or between-participant comparisons of evaluations of

the same object, with between-participant comparisons showing lower reliability estimates (.30–.75) compared to within-participant comparisons (~.80).

- d** Reliability estimates tend to be lower (.40–.60) for second and subsequent IATs if more than one IAT is administered in the same session.

Semantic Priming Tasks

A somewhat less common, though very similar, paradigm is Wittenbrink, Judd, and Park's (1997) semantic priming task. The basic procedure of this measure is analogous to Fazio et al.'s (1986) evaluative priming task, the only difference being that (a) participants are presented with meaningful words and meaningless letter strings as target stimuli and (b) participants' task is to determine as quickly as possible whether the letter string is a meaningful word or a meaningless non-word (*lexical decision task*). To the extent that the presentation of a given prime stimulus facilitates quick responses to a meaningful target word (compared to a baseline prime), the prime stimulus is assumed to be associated with the semantic meaning of the target word. For example, in an application of the task to racial stereotypes, Wittenbrink *et al.* (1997) found facilitated responses to trait words related to the stereotype of African Americans (e.g., athletic, hostile) when participants were primed with the word "black" before the presentation of the target words. In contrast to Fazio et al.'s (1986) evaluative priming task, Wittenbrink et al.'s (1997) paradigm is primarily concerned with semantic associations between a target object and a semantic concept (e.g., associations between *self* and *extraverted*) rather than evaluative associations between a target object and its valence (e.g., associations between *self* and *positive*).

Another variant of semantic priming that is procedurally closer to Fazio et al.'s (1986) evaluative priming task includes only meaningful words as target stimuli, with participants being asked to categorize the target words in terms of their semantic rather than evaluative meaning (*semantic decision task*). For example, Banaji and Hardin (1996) presented participants with prime words referring to stereotypically male or stereotypically female occupations (e.g., nurse, doctor), which were followed by male or female pronouns (e.g., he, she). Participants' task was to classify the pronouns as male or female as quickly as possible. Results showed that participants were faster in responding to the male and female pronouns on stereotype-compatible trials (e.g., nurse-she, doctor-he) than stereotype-incompatible trials (e.g., nurse-he, doctor-she). An important difference between the two kinds of priming tasks is that lexical classifications

(i.e., word vs. non-word) tend to be substantially faster than evaluative or semantic classifications, which leads to smaller effect sizes in priming tasks using lexical classifications. Because priming effects on lexical classifications are often in the range of only a few milliseconds, they are particularly prone to measurement error (e.g., caused by distraction), which poses a challenge to the reliability of semantic priming paradigms using lexical decision tasks.

Affect Misattribution Procedure

A relatively recent but already very popular measure is Payne, Cheng, Govorun, and Stewart's (2005) affect misattribution procedure (AMP). In this task, participants are briefly presented with a prime stimulus, which is followed by a brief presentation of a neutral Chinese ideograph. The Chinese ideograph is then replaced by a black-and-white pattern mask, and participants' task is to indicate whether they consider the Chinese ideograph as visually more pleasant or visually less pleasant than the average Chinese ideograph. The typical finding is that the neutral Chinese ideographs tend to be evaluated more favorably when participants have been primed with a positive stimulus than when they have been primed with a negative stimulus. Although responses in the AMP may seem rather easy to control, priming effects in the AMP have been shown to emerge even when participants are instructed not to let the prime stimuli influence the evaluation of the ideographs and even when they were given detailed information about how the prime stimuli may influence their responses on the task (Payne et al., 2005).

As with Fazio et al.'s (1986) evaluative priming task, the AMP can be used to assess evaluative responses toward any kind of stimuli that can be used as primes in the task. Yet, a major advantage of the AMP is that it shows higher effect sizes and reliability estimates that are comparable to the ones of traditional self-report measures (see Table 12.2). Combined with the procedural advantages of sequential priming (e.g., compatible and incompatible trials being intermixed rather than blocked), these features make the AMP one of the most promising alternatives to the IAT to date. Recently, researchers have also started to investigate the usefulness of the AMP to measure semantic associations, which broadens its potential applicability. For example, using a modified version of the AMP, Gawronski and Ye (2011) asked participants to guess whether the Chinese ideographs referred to a male or a female name. As primes they used words referring to stereotypically male occupations (e.g., doctor) or stereotypically female occupations (e.g., nurse). Results showed that participants were more

likely to guess “male” than “female” when they were primed with a stereotypically male occupation than when they were primed with a stereotypically female occupation. Moreover, priming scores were systematically related to self-report measures of hostile and benevolent sexism (Glick & Fiske, 1996) but not perceptions of gender discrimination, suggesting that the priming effects resulting from gender-stereotypical occupations are genuinely related to the endorsement of sexist attitudes instead of reflecting mere knowledge of unequal gender distributions in these occupations. A noteworthy caveat is that participants may sometimes base their responses on intentional evaluations of the prime stimuli instead of the neutral Chinese ideographs, which could potentially undermine the implicit nature of the task (Bar-Anan & Nosek, 2012). However, intentional ratings seem less prevalent in light of recent findings by Payne, Brown-Iannuzzi, Burkley, Arbuckle, Cooley, Cameron, & Lundberg (2013), who showed that relations between AMP effects and self-reported use of the prime stimuli tend to reflect retrospective confabulations rather than causal effects of intentions during the evaluation of the Chinese ideographs .

Go/No-Go Association Task

Nosek and Banaji's (2001) go/no-go association task (GNAT) was inspired by the basic structure of the IAT with an attempt to make the task amenable for the assessment of associations involving a single target concept (e.g., evaluations of black people) rather than two target concepts (e.g., relative preferences for white over black people). Toward this end, participants are asked to show a *go* response to different kinds of target stimuli (e.g., by pressing the space bar) and a *no-go* response to distracter stimuli (i.e., no button press). In one block of the task, the targets include stimuli related to the target concept of interest (e.g., black faces) and stimuli related to one pole of a given attribute dimension (e.g., positive words); the distracters typically include stimuli related to the other pole of the attribute dimension (e.g., negative words). In a second block, the classification of the particular attribute poles as targets and distracters is reversed (e.g., *go* for black faces and negative words, and *no-go* for positive words). GNAT trials typically include a response deadline, such that participants are asked to show a *go* response to the targets before the expiration of that deadline (e.g., 600 msec). Error rates are analyzed by means of signal detection theory (Green & Swets, 1966), such that differences in sensitivity scores (d') between the two pairings of *go* trials (e.g., black-positive vs. black-negative) are interpreted as an index of associations between the target concept of interest and

the respective attributes. Like the IAT, the GNAT is quite flexible in its application, in that targets and distracters may include a variety of concepts and attributes, including evaluative and semantic attributes associated with individuals, groups, and nonsocial objects (e.g., partner evaluations, self-concept, racial prejudice, consumer preferences). The average reliability of the GNAT is lower compared to the Single Category IAT and the AMP, but still higher compared to the evaluative priming task (see [Table 12.2](#)). A potential problem of the GNAT is that it retains the original block structure of the IAT, which has been linked to various sources of systematic measurement error (Teige-Mocigemba et al., [2010](#)).

Extrinsic Affective Simon Task

Another measure that has been designed to resolve procedural limitations of the IAT is the extrinsic affective Simon task (EAST; De Houwer, [2003](#)). In the critical block of the task, participants are presented with target words (e.g., “beer”) that are shown in two different colors (e.g., yellow vs. blue) and with positive and negative words that are shown in white color. Participants are instructed to categorize the presented words in terms of their valence when they are shown in white color and to categorize them in terms of their color when they are colored. For example, in an EAST designed to measure evaluations of alcoholic beverages, participants may be presented with positive and negative words in white (e.g., spider, sunrise) and with names of alcoholic and nonalcoholic beverages (e.g., beer, soda) that are presented in yellow on some trials and in blue on others. Participants’ task is to press a left-hand key when they see a white word of negative valence or a word printed in blue and to press a right-hand key when they see a white word of positive valence or a word printed in yellow. To the extent that participants show faster (or more accurate) responses to a colored word when the required response to this word is combined with a positive as compared to a negative response, it is inferred that participants showed a positive response to the object depicted by the colored word. Although the EAST was originally designed as a measure of evaluative responses, a number of studies have demonstrated its applicability to other domains, such as the assessment of self-related associations (e.g., Teige, Schnabel, Banse, & Asendorpf, [2004](#)).

A typical EAST includes a total of three blocks: two practice blocks and one critical block. In the first block, participants are presented with the colored target words, which have to be categorized in terms of their color. In the second block,

participants are presented with positive and negative words in white, which have to be categorized in terms of their valence. In the critical third block, the two categorization tasks are combined, such that participants are presented with white and colored words in alternating order. Participants' task is to categorize the words in terms of their valence when they are presented in white and to categorize the words in terms of their color if they are colored.

Although the EAST resolves many of the procedural limitations of the IAT, its average reliability is less than satisfying (see [Table 12.2](#)). De Houwer and De Bruycker (2007) speculated that the low reliability of the EAST stems from the fact that participants do not have to process the meaning of the colored target stimuli for the color-based responses in the task. To overcome this limitation, they developed a modified version of the EAST in which participants are forced to process the meaning of the target stimuli. The identification EAST (ID-EAST) includes presentations of target and attribute words in uppercase and lowercase letters. Positive and negative attribute words have to be categorized in terms of their valence irrespective of whether they are displayed in uppercase or lowercase; the target words have to be categorized depending on whether they are presented in uppercase or lowercase. For example, in an ID-EAST designed to measure evaluative responses to beer, participants may be presented with positive and negative words and the word "beer" in either uppercase or lowercase. Participants' task would be to categorize the attribute words in terms of their valence by pressing one of two response keys. However, for the word "beer," participants would be instructed to press one response key when it is presented in uppercase letters and the opposite key when it is presented in lowercase letters. Because the attribute words are also presented in uppercase and lowercase, participants are therefore required to process the semantic meaning of the word "beer" before they are able to identify the correct response key. This procedural modification increased the reliability of the EAST, although it is still somewhat lower than the average reliabilities of the IAT and the AMP (see [Table 12.2](#)).

Approach-Avoidance Tasks

Another set of paradigms can be subsumed under the general label *approach-avoidance tasks*. The general assumption underlying these tasks is that positive stimuli facilitate approach reactions and inhibit avoidance reactions, whereas negative stimuli facilitate avoidance reactions and inhibit approach reactions. In the first empirical demonstration of such effects, Solarz (1960) found that

participants were faster pulling a lever toward them (approach) in response to positive compared to negative words. Conversely, participants were faster pushing a lever away from them (avoidance) in response to negative compared to positive words. Expanding on these findings, Chen and Bargh (1999) showed that these effects emerge even if the required response is unrelated to the valence of the stimuli (e.g., approach as soon as a word appears on the screen vs. avoid as soon as a word appears on the screen). However, in contrast to earlier interpretations of these effects as resulting from direct, inflexible links between motivational orientations and particular motor actions (contraction of flexor muscle = approach; contraction of extensor muscle = avoidance), accumulating evidence suggests that congruency effects in approach-avoidance tasks depend on the evaluative meaning that is assigned to a particular motor action in the task. For example, Eder and Rothermund (2008) found that participants are faster pulling a lever (flexor contraction) in response to positive words and faster pushing a lever (extensor contraction) in response to negative words when the required motor responses were described as pull (i.e., positive meaning attributed to flexor contraction) and push (i.e., negative meaning attributed to extensor contraction). However, these effects were reversed when the same motor responses were described as upward (i.e., positive meaning attributed to extensor contraction) and downward (i.e., negative meaning attributed to flexor contraction). These results indicate that the particular descriptions of the required motor actions can influence the direction of congruency effects in approach-avoidance tasks. Hence, carefully designed instructions with unambiguous response labels are important to avoid misinterpretations of the resulting scores.

Although most studies have used variations of the aforementioned standard paradigm, noteworthy modifications include the Evaluative Movement Assessment (EMA), which includes left-right responses and visual depictions of their respective meanings (Brendl, Markman, & Messner, 2005), and the Implicit Association Procedure (IAP), in which motor movements are used to assess self-related associations (Schnabel, Banse, & Asendorpf, 2006). An important caveat regarding the use of approach-avoidance tasks is that their reliabilities vary substantially as a function of specific task characteristics (see Table 12.2). For example, reliability estimates are lower for tasks in which stimulus valence is response-irrelevant compared with tasks in which stimulus valence is response-relevant (e.g., Field, Caren, Fernie, & De Houwer, 2011; Krieglmeier & Deutsch, 2010). Moreover, reliability estimates for the EMA tend to be lower for between-participant comparisons of evaluations of the same object compared to

within-participant comparisons of preferences for different objects (see [Table 12.2](#)).

Sorting Paired Features Task

A relatively novel procedure is the sorting paired features (SPF) task, which measures four separate associations in a single response block (Bar-Anan, Nosek, & Vianello, 2009). By using combinations of two simultaneously presented stimuli and four (instead of two) response options, the SPF task breaks the four associations that are confounded in the standard IAT into separate indices. For example, in an application of the SPF task to measure racial prejudice, participants may be presented with pairs of faces and words that involve (a) a white face and a positive word, (b) a black face and a positive word, (c) a white face and a negative word, and (d) a black face and a negative word. Participants' task is to press one of four response keys depending on the particular stimulus combination. Across four blocks of the task, the response key assignment is set up in a manner such that one stimulus dimension is mapped along a vertical response dimension (e.g., positive-right, negative-left), whereas the other stimulus dimension is mapped onto a horizontal response dimension (e.g., white-up, black-down). These mappings are counterbalanced across the four blocks, such that each pair of categories is mapped once with each of the four response keys over the course the task.

For example, in a first block of the race SPF task, combinations of white faces and positive words may require a response with the upper right key (e.g., O); combinations of white faces and negative words may require a response with the upper left key (e.g., W); combinations of black faces and positive words may require a response with the lower right key (e.g., C); and combinations of black faces and negative words may require a response with the lower left key (e.g., M). The key assignment for one stimulus dimension may then be switched in the second block, such that stimulus combinations with positive words go to the left and stimulus combinations with negative words got to the right, while keeping the response dimension for the target category constant (i.e., white-up, black-down). The third and fourth block would then use the two valence mappings with the opposite mapping for the target category (i.e., white-down, black-up). An index of the association between two concepts is calculated by subtracting a participant's mean response latency on all trials with the relevant stimulus combination (e.g., white-positive) from this participant's mean latency on all types of trials (e.g., white-positive; white-negative; black-positive; black-

negative), divided by the standard deviation of the participant's response latencies on all trials. In their original presentation of the SPF task, Bar-Anan *et al.* (2009) report internal consistencies (Spearman-Brown) of the four individual scores ranging between .39 and .71 and test-retest reliabilities between .51 and .60. So far, the SPF has been successfully applied to assess race-related associations and associations related to political attitudes (e.g., Democrats vs. Republicans), although additional research seems desirable to corroborate the validity of the task.

Implicit Relational Assessment Procedure

The implicit relational assessment procedure (IRAP) was developed by Barnes-Holmes and colleagues based on their behavior-analytic theory of human language and thinking (for a review, see Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010). On each trial of an IRAP, participants are presented with two stimuli on the screen (e.g., a picture of an overweight person and a positive word), and participants are trained to identify as quickly as possible which of two keys they are required to press in response to a particular stimulus combination. The two response options are labeled to refer to different ways in which the two stimuli might be related (e.g., similar vs. opposite). Typically, participants are faster when the correct response is in line with their beliefs about how the two stimuli are related than when the correct response contradicts their beliefs about the relation between the two stimuli (for details regarding the scoring of IRAP data, see Barnes-Holmes et al., 2010).

For example, participants might be presented with a picture of a slim person and the word “good,” a picture of a slim person and the word “bad,” a picture of an overweight person and the word “good,” or a picture of an overweight person and the word “bad.” Depending on the specific picture-word combination, participants are trained to press either a key that indicates that the picture and the word are similar or a key that indicates that the picture and the word are opposite. Specifically, participants may have to press the “similar” key for slim-good and overweight-bad combinations and the “opposite” key for slim-bad and overweight-good combinations in some blocks of the task. In other blocks, participants may have to press the “similar” key for slim-bad and overweight-good combinations and the “opposite” key for slim-good and overweight-bad combinations. Whereas in the first type of blocks, the relational meaning of the required key responses is compatible with the attitudinal beliefs of those participants who like slim people or dislike overweight people, the relational

meaning in the second type of blocks is compatible with the attitudinal beliefs of participants who like overweight people or dislike slim people. Although the task structure of the IRAP has some resemblance to the IAT, in that it combines associations between two target objects and two attributes, the IRAP has been shown to be amenable to the measurement of attitudes toward individual objects in a nonrelative manner (e.g., Roddy, Stewart, & Barnes-Holmes, 2011).

A unique characteristic of the IRAP is that it is designed to capture propositional beliefs rather than mere associations. Whereas associations link two concepts without specifying the particular way in which these concepts are related, propositional beliefs do specify the way in which concepts are related (Hughes, Barnes-Holmes, & De Houwer, 2011). For example, a person might simultaneously believe that he *is* bad and that he *wants to be* good. An implicit measure that captures mere associations would not be able to differentiate between these two beliefs. Instead, it would show evidence for associative links between *self* and *bad* and, at the same time, between *self* and *good*. In the IRAP, these beliefs can be differentiated by using different types of stimulus combinations (e.g., the expressions *I am* and *I am not* versus the expressions *I want to be* and *I do not want to be* presented in combination with the words “good” and “bad”; Remue, De Houwer, Barnes-Holmes, Vanderhasselt, & De Raedt, in press). Although the IRAP has been primarily used to measure evaluative beliefs (e.g., *being slim is good*), it is also amenable to the assessment of semantic beliefs (e.g., *I am able to approach spiders*; Nicholson & Barnes-Holmes, 2012). Reliability estimates, however, differ substantially between studies, ranging from values as low as .23 to values as high as .81. Although little is known about procedural variables that moderate the reliability of IRAP effects, some studies suggest that the reliability of the IRAP increases with decreases in the response deadline (Barnes-Holmes et al., 2010) .

Action Interference Paradigm

The action interference paradigm (AIP) has been developed for research involving very young children, who might get overwhelmed by the complex task requirements of other paradigms. For example, in one application to study the development of gender stereotypes, Banse, Gawronski, Rebetez, Gutt, and Morton (2010) told young children that Santa Claus needs their help in delivering Christmas presents to other children. In a first block of the task, the children were told that the first family had a boy and a girl and that the boy would like to get trucks and the girl would like to get dolls. The children were

then shown pictures of trucks and dolls on the screen, and they were asked to give the presents to the kids as quickly as possible by pressing the buttons of a response box that were marked with pictures of the boy and the girl. In a second block, the children were told that they are now at the house of another family, which also had a boy and a girl. However, this boy would like to get dolls and the girl would like to get trucks. The children were then shown the same pictures of trucks and dolls, and they were asked to press the response buttons that were marked with the pictures of another boy and girl. Controlling for various procedural features, Banse *et al.* (2010) found that children were faster in making stereotype-compatible assignments (i.e., boy-truck, girl-doll) compared to stereotype-incompatible assignments (i.e., boy-doll, girl-truck), which was interpreted as evidence for spontaneous gender stereotyping in children.

Among the paradigms reviewed in the current chapter, the AIP is the most content-specific measure, in that the original variant is particularly designed for the assessment of gender stereotypes. Nevertheless, it seems possible to modify the AIP for the assessment of other constructs. For example, to assess evaluative responses in the domain of racial prejudice, the gender categories could be replaced by racial categories and the assignment task may involve the distribution of desirable and undesirable objects to black and white children. However, it is important to point out that applications of the AIP to other domains require a different framing of the task in the instructions. In addition, it is worth noting that the internal consistency of the AIP is relatively low, with Cronbach's Alpha values in the range of .30 and .50 (Gawronski et al., 2011).

How to Choose a Measurement Procedure

Given the large number of available paradigms, a common question by novices is which of them they should choose for their own research. In making this choice, we believe that it is important to consider that measurement procedures are tools and different types of research questions require different kinds of tools. Thus, instead of recommending a particular paradigm as the “best” one, we try to provide some heuristics that might be useful in identifying the most suitable paradigm for a particular research question.

A first issue is that the reviewed paradigms differ considerably with regard to their flexibility. Whereas some tasks have been developed to assess either semantic or evaluative representations, others are more specific in the type of questions for which they can be used (see Table 12.2). Thus, a first constraint on the choice of a particular measure is whether one's research question involves

semantic or evaluative representations. Similarly, whereas some measures are suitable to measure representations involving individual targets and individual attributes, other paradigms involve comparisons between pairs of targets and pairs of attributes (see [Table 12.2](#)). Thus, to maximize the conceptual overlap between research design and implicit measurement scores, it is important to consider whether one's research question involves a comparison between pairs of targets and pairs of attributes. For example, the comparative structure of the IAT seems less problematic if one is interested in how gender-stereotypical associations influence impressions of men versus women who engage in stereotype-congruent versus stereotype-incongruent behaviors (e.g., Gawronski, Ehrenberg, Banse, Zukova, & Klauer, 2003). However, the IAT seems less suitable if one is interested in evaluative responses toward a particular target person, which are easier to capture with sequential priming tasks (e.g., Rydell & Gawronski, 2009).

Another important consideration is the wide range of reliability estimates that have been reported for different implicit measures (see [Table 12.2](#)). Whereas some paradigms have consistently shown satisfying reliability estimates across different applications, others suffer from large variations or clearly unsatisfactory psychometric properties. Although concerns about low reliability tend to be more common in personality psychology than in social psychology, low internal consistency can be a problem in both individual difference and experimental designs. On the one hand, low internal consistency can distort the rank order of participants with regard to a particular construct, which reduces correlations to other measures (e.g., in studies on the prediction of behavior). On the other hand, low internal consistency can reduce the probability of identifying effects of experimental manipulations (e.g., in studies on attitude change), which includes both initial demonstrations of an experimental effect and replications of previously obtained effects (LeBel & Paunonen, 2011).³

Finally, it is important to point out that none of the reviewed measures is perfect, and that any choice between these tasks involves a trade-off between desirable and undesirable features. In addition to structural aspects and reliability estimates, examples of other relevant features include the overall length of the task and its suitability for populations that may be less experienced with computer-based tasks than undergraduate students are (e.g., children, older adults). Of course, the relative importance of these features depends on one's research question, which makes it difficult to make strong recommendations on a priori grounds. Nevertheless, we hope that our review and the aforementioned heuristics are helpful in making informed decisions about which measure might

be most useful for a given research question.

What Can We Learn from Implicit Measures?

The number of studies using implicit measures has grown exponentially over the past decade, and their findings have influenced virtually every area of psychology (for an overview, see Gawronski & Payne, 2010). A popular theme in these studies concerns dissociations between explicit and implicit measures. Such dissociations are often interpreted with reference to dual-process theories, in that the different measures are assumed to reflect the operation of distinct mental processes (e.g., Fazio, 2007; Gawronski & Bodenhausen, 2006; Rydell & McConnell, 2006; Strack & Deutsch, 2004). In the following sections, we provide a brief overview of the insights that can be gained from these dissociations with regard to the prediction of behavior, the prediction of biases in information processing, and the formation and change of mental representations.

Implicit Measures as a Tool for Predicting Behavior

Two of the first questions that have been asked about implicit measures were: (1) Do implicit measures predict behavior? (2) Do implicit measures add anything to the prediction of behavior over and above explicit measures? Both questions were soon answered positively, and research quickly moved beyond zero-order and additive relations to investigate the conditions under which explicit and implicit measures predict behavior (for reviews, see Friesen, Hofmann, & Schmitt, 2008; Perugini, Richetin, & Zanna, 2010). Inspired by theorizing on attitude-behavior relations, one of the earliest findings was that implicit measures tend to outperform explicit measures in the prediction of spontaneous behavior (e.g., eye gaze in interracial interactions predicted by implicit measures of racial prejudice), whereas explicit measures tend to outperform implicit measures in the prediction of deliberate behavior (e.g., content of verbal responses in interracial interactions predicted by explicit measures of racial prejudice). This double dissociation has been replicated in a variety of domains with several different measures (e.g., Asendorpf, Banse, & Mücke, 2002; Fazio, Jackson, Dunton, & Williams, 1995).

Expanding on the idea that the predictive validity of implicit and explicit measures is determined by automatic versus controlled features of the to-be-predicted behavior, several recent studies found that a given behavior showed stronger relations to explicit measures compared with implicit measures under

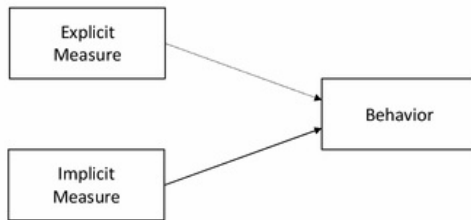
conditions of unconstrained processing resources. Yet, the same behavior showed stronger relations to implicit measures than explicit measures when processing resources were depleted. For example, candy consumption under cognitive depletion has been shown to be related to an implicit measure of candy attitudes, but to an explicit measure of candy attitudes under control conditions (e.g., Hofmann, Rauch, & Gawronski, 2007). Similar findings have been obtained for the motivation to engage in elaborate cognitive processing (e.g., Scarabis, Florack, & Gosejohann, 2006). Adopting an individual difference approach, a number of studies have demonstrated that explicit measures are better predictors of behavior for people with a preference for rational thinking styles, whereas implicit measures are better predictors of behavior for people with a preference for intuitive thinking styles (e.g., Richetin, Perugini, Adjali, & Hurling, 2007).

Deviating from approaches in which implicit and explicit measures are seen as competitors in the prediction of behavior, several studies have investigated interactive relations between the two kinds of measures. The general assumption underlying these studies is that discrepancies between implicit and explicit measures are indicative of an unpleasant psychological state that people aim to reduce (Rydell, McConnell, & Mackie, 2008). For example, people showing large discrepancies on implicit and explicit measures of a particular psychological attribute (e.g., attitude, self-concept) have been shown to elaborate attribute-related information more extensively than people with small discrepancies (e.g., Briñol, Petty, & Wheeler, 2006). In a similar vein, combinations of high self-esteem on explicit measures and low self-esteem on implicit measures have been shown to predict defensive behaviors, such as favoring one's in-group over out-groups and dissonance-related attitude change (e.g., Jordan, Spencer, Zanna, Hoshino-Browne, & Correll, 2003).

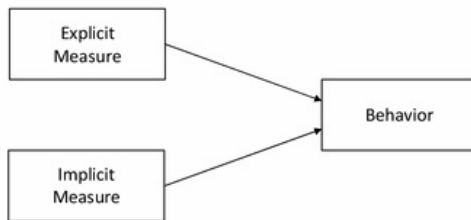
Perugini *et al.* (2010) have provided a conceptual summary of different patterns in the prediction of behavior by implicit measures (see Figure 12.1). These patterns include: (1) single association patterns in which implicit measures, but not explicit measures, predict the relevant behavior; (2) additive patterns in which implicit and explicit measures jointly predict the relevant behavior; (3) double dissociation patterns in which implicit and explicit measures uniquely predict different kinds of behavior; (4) moderation patterns in which implicit and explicit measures predict the relevant behavior under different conditions; and (5) multiplicative patterns in which implicit and explicit measures interactively predict the relevant behavior. All of these patterns have been demonstrated in the literature and they are generally consistent with current

dual-process theorizing (e.g., Fazio, 2007; Strack & Deutsch, 2004). However, their boundary conditions are still not well understood, which makes it difficult to predict particular outcomes in an a priori manner. Thus, an important task for future research is to identify the particular conditions under which each of these patterns occurs.

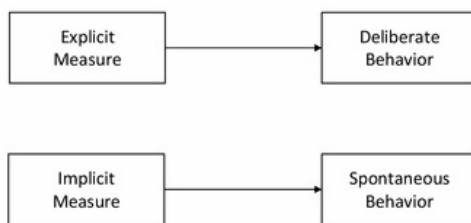
1) Simple Association Pattern



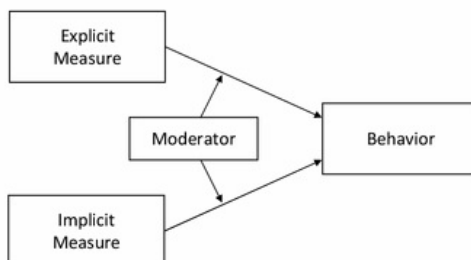
2) Additive Pattern



3) Double Dissociation Pattern



4) Moderation Pattern



5) Multiplicative Pattern

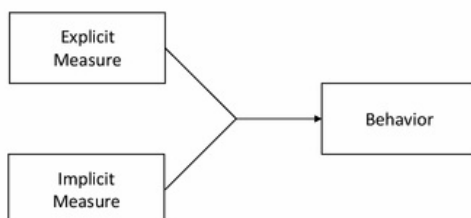


Figure 12.1. Patterns of behavior prediction by implicit measures. Figure adapted from Perugini, Richetin, and Zogmaister (2010). Reprinted with permission.

Implicit Measures as a Tool for Predicting Biases in Information Processing

Although double dissociation patterns in the prediction of spontaneous and deliberate behavior are well established in the literature, there are several studies in which implicit measures outperformed explicit measures in the prediction of deliberate judgments, even when there is evidence for the construct validity of the explicit measure. These findings suggest that the representations captured by implicit measures may bias the processing of available information, which can influence deliberate judgments that are based on this information. One example in this regard is the interpretation of ambiguous information. Previous research has shown that contextual cues can distort the interpretation of ambiguous information in a manner that is consistent with the subjective meaning of the contextual cues. For example, in the domain of racial prejudice, the same ambiguous behavior is often interpreted in a positive manner when the actor is white, but negatively when the actor is black (e.g., Sagar & Schofield, 1980). Although self-reported interpretations of ambiguous behavior may be regarded as an example of deliberate behavior, interpretational biases have been found to reveal stronger relations to implicit measures compared with explicit measures (e.g., Gawronski, Geschke, & Banse, 2003; Hugenberg & Bodenhausen, 2003). This asymmetry has been interpreted as evidence that biases in the interpretation of ambiguous information are driven by the associations that are automatically activated by contextual cues rather than by perceivers' explicitly held beliefs.

Another example of bias in information processing is selective information search. A common finding in the literature on cognitive dissonance is that people selectively expose themselves to information that is consistent with their self-reported attitudes (for a meta-analysis, see Hart, Albarracín, Eagly, Brechan, Lindberg, & Merrill, 2009). Although this bias has been shown to be reduced for attitudes that are not held with conviction, research using implicit measures has found that even undecided individuals have a tendency to selectively expose themselves to particular information (Galdi, Gawronski, Arcuri, & Fries, 2012). Whereas selective exposure in decided participants showed stronger relations to explicit than implicit measures, selective exposure in undecided

individuals showed stronger relations to implicit than explicit measures. Such biases in information processing explain why implicit measures are capable of predicting future choices and decisions that seem highly deliberate, such as voting behavior and other political decisions (e.g., Galdi, Arcuri, & Gawronski, 2008 ; Payne, Krosnick, PASEK, Lelkes, Akhtar, & Thompson, 2010). For example, undecided voters may selectively expose themselves to information that is consistent with their implicit preference, and this biased set of information may ultimately provide the basis for their deliberate decision to vote for a particular candidate. Thus, to the extent that deliberate choices are based on the information that is available to an individual and the representations captured by implicit measures predict processing biases in the acquisition of this information (e.g., biased interpretation, selective exposure), implicit measures can be expected to make a unique contribution to the prediction of future decisions even when these decisions are highly deliberate.

Implicit Measures as a Tool for Understanding the Formation and Change of Mental Representations

Given the available evidence for dissociations in studies using implicit and explicit measures as predictor variables, an interesting question concerns potential dissociations when implicit and explicit measures are used as dependent variables. This question has been particularly dominant in research on attitude formation and change, which has shown various dissociations in the antecedents of attitudes captured by implicit and explicit measures (for a review, see Gawronski & Bodenhausen, 2006). Whereas some studies found effects on explicit measures but not on implicit measures (e.g., Gregg, Seibt, & Banaji, 2006), others showed effects on implicit measures but not explicit measures (e.g., Olson & Fazio, 2006). Other studies, however, found corresponding effects on both explicit and implicit measures (e.g., Whitfield & Jordan, 2009). These inconsistent patterns posed a challenge to traditional theories of attitude formation and change, which inspired the development of novel theories that have been designed to explain potential dissociations between implicit and explicit measures of attitudes (e.g., Gawronski & Bodenhausen, 2006; Petty, Briñol, & DeMarree, 2007; Rydell & McConnell, 2006).

One example is Gawronski and Bodenhausen's (2006, 2011) associative-propositional evaluation (APE) model, which distinguishes between the activation of associations in memory (*associative process*) and the validation of momentarily activated information (*propositional process*). According to the

APE model, processes of association activation are driven by principles of similarity and contiguity; processes of propositional validation are assumed to be guided by principles of logical consistency. This distinction between associative and propositional processes is further linked to implicit and explicit measures, in that implicit measures are assumed to reflect the behavioral outcome of associative processes, whereas explicit measures are assumed to reflect the behavioral outcome of propositional processes. Drawing on several assumptions about mutual interactions between associative and propositional processes, the APE model has generated a number of novel predictions regarding the conditions under which a given factor should lead to (a) changes on explicit but not implicit measures; (b) changes on implicit but not explicit measures; (c) corresponding changes on explicit and implicit measures, with changes on implicit measures being mediated by changes on explicit measures; and (d) corresponding changes on explicit and implicit measures, with changes on explicit measures being mediated by changes on implicit measures. For example, consistent with the predictions of the APE model, cognitive dissonance has been shown to change explicit, but not implicit, evaluations (e.g., Gawronski & Strack, 2004). Conversely, repeated pairings of a neutral conditioned stimulus (CS) with a valenced unconditioned stimulus (US) have been shown to change implicit evaluations of the CS, whereas explicit evaluations were affected only when participants were instructed to introspect on their gut feelings toward the CS (e.g., Gawronski & LeBel, 2008). Although the APE model is just one among several theories that aim to account for dissociations in the antecedents of implicit and explicit measures (e.g., Petty et al., 2007; Rydell & McConnell, 2006), research including both kinds of measures as dependent variables can help provide deeper insights into the formation and change of mental representations.

Some Caveats Regarding the Interpretation of Implicit Measures

As we outlined in the preceding section, implicit measures have provided important insights into the determinants of behavior, biases in information processing, and the formation and change of mental representations. At the same time, there are a number of misconceptions about the type of information implicit measures can provide (Gawronski, 2009). In the current section, we discuss several assumptions that are quite common in the interpretation of implicit measures, yet questionable on the basis of the available evidence.

Conscious versus Unconscious Representations

A very common assumption is that indirect measurement procedures provide a window into unconscious representations, whereas direct self-report measures reflect conscious representations (e.g., Greenwald & Banaji, 1995). The central idea underlying this assumption is that self-report measures require introspective access to the to-be-measured memory contents, which undermines their suitability for the measurement of memory contents that are unconscious. In contrast, indirect measures do not presuppose introspective access for the measurement of memory contents, which makes them amenable for the assessment of unconscious memory contents. It is important to note that any such claims represent empirical hypotheses that have to be tested as such. To be sure, it is true that indirect measures do not require introspective access for the assessment of memory contents. However, this does not imply that the memory contents that are assessed by these measures are indeed unconscious.

A common argument in support of the unconsciousness claim is that the two types of measures often show rather low correlations. Of course, if the memory contents captured by an indirect measure are unconscious, their correspondence to self-report measures may be low. However, dissociations between different measures can be the result of multiple other factors that do not imply lack of introspective access (for a review, see Hofmann, Gschwendner, Nosek, & Schmitt, 2005). For example, research on prejudice has shown that correlations between self-report and evaluative priming measures are higher when participants' motivation to control prejudiced reactions is low than when it is high (e.g., Dunton & Fazio, 1997), and the same effects have been shown for the IAT (e.g., Gawronski et al., 2003). Moreover, several studies in the domain of attitudes have shown that correlations between the two kinds of measures are higher when participants focus on their gut feelings toward the attitude object (e.g., Gawronski & LeBel, 2008). Taken together, these results suggest that low correspondence between direct measures and indirect measures may not result from a lack of introspective access to the memory contents captured by the latter type of measure. Instead, their correspondence may be determined by a variety of other factors, such as motivational influences and introspective mindsets during judgment. Thus, interpretations of the two kinds of measures as providing access to conscious versus unconscious representations are difficult to reconcile with the available evidence (for a review, see Gawronski, Hofmann, & Wilbur, 2006).

Old versus New Representations

Another common assumption is that implicit measures reflect highly stable, old representations that have not been replaced by more recently acquired, new representations. The central idea underlying this assumption is that previously formed representations are not erased from memory when people acquire new information that is inconsistent with these representations. To the extent that earlier acquired knowledge is often highly overlearned, older representations are assumed to be activated automatically upon encountering a relevant stimulus. In contrast, more recently acquired knowledge is usually less well learned, which implies that newer representations require controlled processes to be retrieved from memory. With regard to attitudes, for example, it is often assumed that people can have two distinct attitudes toward the same object: an earlier acquired “implicit” attitude that is activated automatically upon encountering a relevant stimulus, and a more recently acquired “explicit” attitude that requires conscious effort to be retrieved from memory (e.g., Wilson, Lindsey, & Schooler, 2000). This distinction between (old) implicit and (new) explicit representations is often mapped onto particular kinds of measures, such that indirect measures are assumed to tap into earlier, acquired implicit representations, whereas direct self-report measures are claimed to capture more recently acquired, explicit representations (e.g., Rudman, 2004).

As with interpretations in terms of conscious versus unconscious representations, the claim that different kinds of measurement procedures are differentially sensitive to old versus newly formed representations is an empirical hypothesis that needs to be verified with relevant data. Consistent with this claim, there is some evidence showing an impact of recent experiences on explicit, but not implicit, measures (e.g., Gawronski & Strack, 2004; Gregg et al., 2006). However, there is also a large body of research showing the opposite pattern (e.g., Gawronski & LeBel, 2008; Olson & Fazio, 2006). The latter findings are difficult to reconcile with claims that implicit measures tap into highly overlearned, old representations, and that explicit measures reflect recently acquired, new representations.

Dissociations between Explicit and Implicit Measures

Implicit measures become particularly interesting when they show dissociations with explicit measures. However, when interpreting such dissociations, it is important to consider a number of potential confounds that may hamper straightforward interpretations of the obtained results. One of the most common

confounds is a mismatch in the relevant target object. For example, researchers interested in racial prejudice often use the race IAT as a measure of implicit prejudice and the Modern Racism Scale (McConahay, 1986) as a measure of explicit prejudice. However, dissociations between the two measures may not necessarily reflect two discrepant racial attitudes, given that the two measures assess evaluative responses to different kinds of objects. Whereas the race IAT captures evaluative responses to black and white faces, the Modern Racism Scale measures perceptions of racial discrimination and evaluative responses to antidiscrimination policies. This concern echoes Ajzen and Fishbein's (1977) correspondence principle in attitude-behavior relations, according to which measures of attitudes and behavior should match with regard to the relevant attitude object. In fact, correlations between implicit and explicit measures are considerably higher when their respective contents match than when their contents mismatch (Hofmann, Gawronski et al., 2005).

In addition to content-related confounds, dissociations between implicit and explicit measures may also be the result of structural task differences (Payne, Burkely, & Stokes, 2008). For example, whereas explicit measures are typically based on participants' responses on rating scales, most implicit measures are based on response latencies or error rates. Hence, even if the two kinds of measures match with regard to their content (e.g., responses to black and white faces), dissociations could also be attributable to differences in the particular aspects of participants' responses that are used to derive the relevant measurement scores (e.g., ratings vs. latencies). To overcome this limitation, Payne *et al.* (2008) presented an extended variant of the AMP that increases the structural fit between implicit and explicit measures of the same construct. The basic structure of the task is similar to Payne et al.'s (2005) original AMP. Yet, the measure is administered in two different ways: an indirect variant for the assessment of implicit measurement outcomes and a direct variant for the assessment of explicit measurement outcomes. Whereas in the indirect variant participants are asked to evaluate the neutral Chinese ideographs and to ignore the prime stimuli, the direct variant asks participants to evaluate the prime stimuli and to ignore the Chinese ideographs. Thus, the two tasks provide measurement outcomes that are comparable not only with regard to the relevant target object (e.g., black and white faces), but also with regard to basic structural features, such as the presentation format and the nature of the relevant responses. Although the two AMP variants showed meaningful differences that are compatible with current theorizing about implicit measures (e.g., the relation between explicit and implicit prejudice scores being moderated by motivation to

control prejudiced reactions), their zero-order correlation was substantially higher compared to the low correlation that is typically found when there is a structural misfit between measures.

Reliability also has to be considered when interpreting dissociations between implicit and explicit measures. Whereas some implicit measures consistently show reliability estimates that are comparable to the ones revealed by explicit measures, others suffer from relatively low reliabilities (see [Table 12.2](#)). Thus, dissociations between implicit and explicit measures may sometimes be attributable to large proportions of measurement error in the implicit measure. Consistent with this concern, Cunningham, Preacher, and Banaji (2001) showed that correlations between implicit and explicit measures are considerably higher when measurement error is taken into account. Because low reliability can also reduce the probability of identifying effects of experimental manipulations (LeBel & Paunonen, 2011), the same concerns apply to studies that compare the relative impact of a given factor on implicit and explicit measures .

Social Desirability, Faking, and Lie Detection

A common assumption in research using implicit measures is that they resolve the well-known problems of social desirability. This assumption is based on the premise that responses on indirect measurement procedures are more difficult to control than are responses on direct measurement procedures. However, several issues have to be considered in this context.

First, it is certainly possible to use implicit measures to rule out social desirability as an alternative explanation for effects obtained with explicit measures. To the extent that both measures show the same effects, it seems rather unlikely that the pattern revealed by the explicit measure is driven by social desirability. However, it is important to note that dissociations between implicit and explicit measures do not necessarily reflect an influence of social desirability on the explicit measure. As we argued earlier in this chapter, dissociations between the two kinds of measures can be attributable to multiple factors over and above social desirability (for a review, see Hofmann, Gschwendner et al., 2005).

Second, it is important to note that responses on indirect measurement procedures are not entirely immune to faking. Although intentional distortions tend to be more difficult on indirect measures compared with direct measures, there is evidence that responses on indirect measures are susceptible to strategic influences to a certain extent (e.g., Klauer & Teige-Mocigemba, 2007; Steffens,

2004).

Third, even if responses on indirect measurement procedures were entirely immune to faking, this does not mean that their measurement outcomes could be used as a lie detector (e.g., Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 2008). To illustrate, consider the use of implicit measures of child-sex associations to identify convicted child molesters (e.g., Gray, Brown, MacCulloch, Smith, & Snowden, 2005). Several studies found that implicit measures are indeed successful in discriminating between pedophiles and non-pedophiles. However, child-sex associations may have their roots in a number of factors other than pedophilia – for example, when a person has been the target of sexual abuse as a child. Because implicit measures are typically unable to distinguish between different sources of mental representations, claims that implicit measures could be used as a lie detector should be treated with caution.

Context Effects

Another common assumption about implicit measures is that they can help researchers resolve the problem of context effects on self-reports. Research on response processes in self-report measures has identified a wide range of contextual factors that can undermine accurate assessment of psychological attributes (for reviews, see Schwarz, 1999; Visser, Krosnick, Lavrakas, & Kim, Chapter 16 in this volume). With the development of indirect measurement procedures that do not rely on self-assessments, many researchers expected to gain direct access to people's “true” personal characteristics without contamination by contextual factors. However, the available evidence suggests that implicit measures are at least as susceptible to contextual influences as explicit measures are (for a review, see Gawronski & Sritharan, 2010). For example, several studies using implicit measures have shown that responses to the same person (e.g., racial minority member) can vary as a function of the context (e.g., family barbeque vs. graffiti wall) in which this person is presented (e.g., Wittenbrink, Judd, & Park, 2001).

Some researchers interpreted these findings as evidence that responses on any type of psychological measure, be it direct or indirect, do not reflect stable trait-like characteristics, but instead are constructed on the spot on the basis of momentarily accessible information (e.g., Schwarz, 2007). Other researchers have argued that contextual influences do not reflect a change in the response to a given object, but rather a change of the target object itself (e.g., Fazio, 2007). For example, evaluative responses to Michael Jordan may differ depending on

whether he is categorized as an athlete or an African American, and momentarily available context cues (e.g., basketball court vs. graffiti wall) may influence how he is categorized in the first place (e.g., Mitchell, Nosek, & Banaji, 2003). According to this view, the relevant category representations may be highly stable, although contextual factors may influence which category representation becomes relevant in a given context. A third class of models takes a position in-between the two opposing camps, arguing that the same object may activate different patterns of stored associations in memory depending on the context in which the object is encountered (e.g., Gawronski & Bodenhausen, 2006). Drawing on the concept of pattern matching in memory retrieval, which associations are activated in a given situation is assumed to depend on the match between momentary input stimuli and the existing structure of associations in memory. Although it is rather difficult to distinguish among the three accounts on the basis of the currently available evidence, the bottom line is that implicit measures are highly sensitive to contextual influences, which challenges the idea that implicit measures provide context-independent assessments of people's "true" representations.

“Automatic” Effects of Experimental Manipulations

A common assumption underlying the use of implicit measures is that the to-be-measured psychological attribute influences measurement outcomes automatically (cf. De Houwer et al., 2009). Based on this assumption, implicit measures are sometimes included as dependent measures in experimental studies to test whether the employed manipulation influences a particular psychological attribute in an automatic fashion. However, such interpretations conflate the impact of the psychological attribute on measurement outcomes with the impact of the experimental manipulation on the psychological attribute (see Figure 12.2). Although such conflations are relatively common in the literature, they are not justified. After all, the implicitness of a given measure speaks only to the automaticity of the impact of the to-be-measured psychological attribute on the measurement outcome (Path B in Figure 12.2); it does not speak to the effect of an experimental manipulation on the psychological attribute (Path A in Figure 12.2).

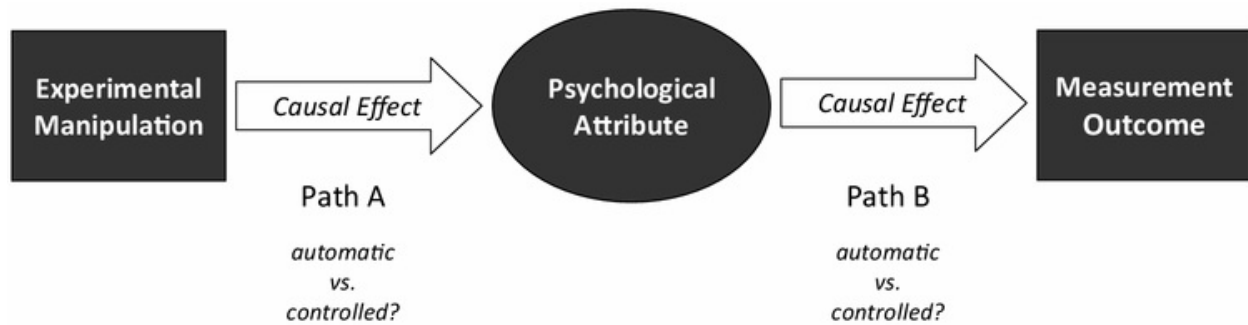


Figure 12.2. Automatic versus controlled effects of an experimental manipulation on a psychological attribute (Path A) and automatic versus controlled effects of a psychological attribute on measurement outcomes (Path B). Empirical evidence for the automatic nature of Path B does not speak to the automatic nature of Path A.

To illustrate this issue, consider a study by Peters and Gawronski (2011) in which participants were asked to recall past behaviors reflecting either extraversion or introversion, and then to complete an IAT designed to measure associations between the self and extraversion/introversion. Results showed that IAT scores of self-extraversion associations were higher when participants were asked to recall extraverted behaviors than when they were asked to recall introverted behaviors. At first glance, one might be tempted to conclude that recalling past behaviors influenced self-representations in an automatic fashion. However, the task of recalling past behaviors was fully conscious, intentional, and controllable, which implies that the experimental manipulation influenced self-representations in a nonautomatic fashion. Of course, it is certainly possible that other experimental manipulations may influence self-representations unconsciously, unintentionally, and uncontrollably. This possibility, however, does not allow one to draw the reverse conclusion that implicit measures can be used to demonstrate the automatic nature of an experimental effect. For example, increased levels of self-esteem on the IAT as a result of personal threat do not necessarily indicate that threat defense mechanisms operate automatically (e.g., Rudman, Dohn, & Fairchild, 2007). Such inferences require additional manipulations, for example, the use of a cognitive load task to investigate the resource (in)dependence of threat defense.

Absolute versus Relative Interpretations

Another important issue concerns metric interpretations of implicit measurement scores. Many of the scoring procedures for implicit measures involve the

calculation of difference scores, in which latencies or error rates on “compatible” trials are compared with the latencies or error rates on “incompatible” trials (or neutral baseline trials). The resulting numerical values are often used to infer a psychological attribute on one side of a continuum if the resulting score is higher than zero (e.g., preference for whites over blacks) and a psychological attribute on the other side of a continuum if the score is lower than zero (e.g., preference for blacks over whites), with a value of zero being interpreted as a neutral reference point. Although metric interpretations of this kind are rather common in the literature, we consider them as problematic for at least two reasons. Aside from the fact that the metric of any given measure remains ambiguous without proper calibration (Blanton & Jaccard, 2006), contingent features of the employed stimulus materials have been shown to influence both the size and the direction of implicit measurement scores (e.g., Bluemke & Fries, 2006; Scherer & Lambert, 2009). Because it is virtually impossible to quantify the contribution of material effects, absolute interpretations of implicit measurement scores are therefore not feasible regardless of whether they involve characteristics of individual participants (e.g., participant X shows a preference for whites over blacks) or samples (e.g., 80% of the sample showed a preference for whites over blacks).

It is important to note that most research questions in social and personality psychology do not require absolute interpretations, but instead are based on relative interpretations of measurement scores. The latter applies to experimental designs in which measurement scores are compared across different groups (e.g., participants in the experimental group show higher scores compared to participants in the control group) as well as individual difference designs in which measurement scores are compared across different participants (e.g., participant A has a higher score compared to participant B). Hence, the aforementioned problems do not necessarily undermine the usefulness of implicit measures in social and personality psychology, although they do prohibit absolute interpretations of measurement scores of individual participants or samples.

Multiple Processes Underlying Implicit Measures

A final caveat concerns the lack of process purity of implicit measures. It is commonly assumed that implicit measures provide direct access to mental associations that are activated automatically upon the encountering a relevant stimulus. However, responses on indirect measurement procedures are the

product of multiple distinct processes that jointly influence performance on the task. To overcome this problem, researchers have developed mathematical modeling techniques that provide a more fine-grained analysis of data obtained with indirect measurement procedures (e.g., Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007; Payne, 2008). The main advantage of these modeling techniques is that they allow researchers to quantify the individual contributions of multiple distinct processes to task-performance instead of relying on a single measurement score. Because the mathematical underpinnings of these procedures go beyond the scope of this chapter, we limit our discussion to a brief description of Conrey et al.'s (2005) quad-model to illustrate how responses on indirect measures depend on multiple processes.⁴

To illustrate the basic assumptions of Conrey et al.'s (2005) quad-model, consider a race IAT with the target categories *black* versus *white* and the attribute categories *pleasant* versus *unpleasant*. In the combined blocks of this IAT, a black face may elicit a response tendency to press the *black* key, and, to the extent that negative associations are activated, another response tendency to press the *unpleasant* key. If *black* and *negative* responses are mapped onto the same key (“compatible” block), responses will be facilitated. If, however, *black* and *negative* responses are mapped onto different keys (“incompatible” block), the tendency to press the *negative* key has to be inhibited so that the accurate tendency to press the *black* key can be executed. Importantly, because the inhibition of the incorrect response tendency requires executive control processes, the impact of race-related associations is confounded with executive control processes in the traditional calculation of IAT scores.

To address this limitation, Conrey et al.'s (2005) quad-model includes statistical parameters for four qualitatively distinct processes: (1) the likelihood that an association-related response tendency is activated (*Association Activation* or *AC*); (2) the likelihood that the correct response to the stimulus can be determined (*Discriminability* or *D*); (3) the likelihood that an automatic association is successfully overcome in favor of the correct response (*Overcoming Bias* or *OB*); and (4) the likelihood that a general response bias (e.g., right-hand bias) drives the response (*Guessing* or *G*).

The proposed interplay of these processes in the quad-model can be depicted as a processing tree that specifies how their joint operation can lead to correct or incorrect responses on compatible and incompatible trials (see [Figure 12.3](#)).⁵ To illustrate the logic of this processing tree, consider the presentation of a black

face in the two combined blocks of the race IAT. If the black face activates a prejudicial response tendency (AC) and participants are able to identify the correct response (D), whether or not the prejudicial response tendency will drive the final response depends on whether participants are able to inhibit the prejudicial response tendency. If they are able to inhibit the prejudicial response tendency (OB), they will show the correct response on both compatible and incompatible trials and regardless of whether the required response is on the left or on the right (first row in [Figure 12.3](#)). However, if they are unable to inhibit the prejudicial response tendency ($1 - OB$), they will show the correct response on compatible trials but an incorrect response on incompatible trials (second row in [Figure 12.3](#)). Moreover, if a prejudicial response tendency is activated (AC) and, at the same time, participants are not able to identify the correct response ($1 - D$), the quad-model assumes that the prejudicial response tendency will drive the final response in the task. In this case, participants will show the correct response on compatible trials but an incorrect response on incompatible trials (third row in [Figure 12.3](#)). If no prejudicial response tendency is activated ($1 - AC$) and participants are able to identify the correct response (D), they will show the correct response on both compatible and incompatible trials and regardless of whether the required response is on the left or on the right (fourth row in [Figure 12.3](#)). Finally, if no prejudicial response tendency is activated ($1 - AC$) and participants are unable to identify the correct response (D), a guessing bias is assumed to drive the final response. For example, if participants show a bias toward responding with the right key (G), they will show the correct response on both compatible and incompatible trials when the correct response is on the right but not when it is on the left (fifth row in [Figure 12.3](#)). Conversely, if participants show a bias toward responding with the left key ($1 - G$), they will show the correct response when the correct response is on the left but not when it is on the right (sixth row in [Figure 12.3](#)).

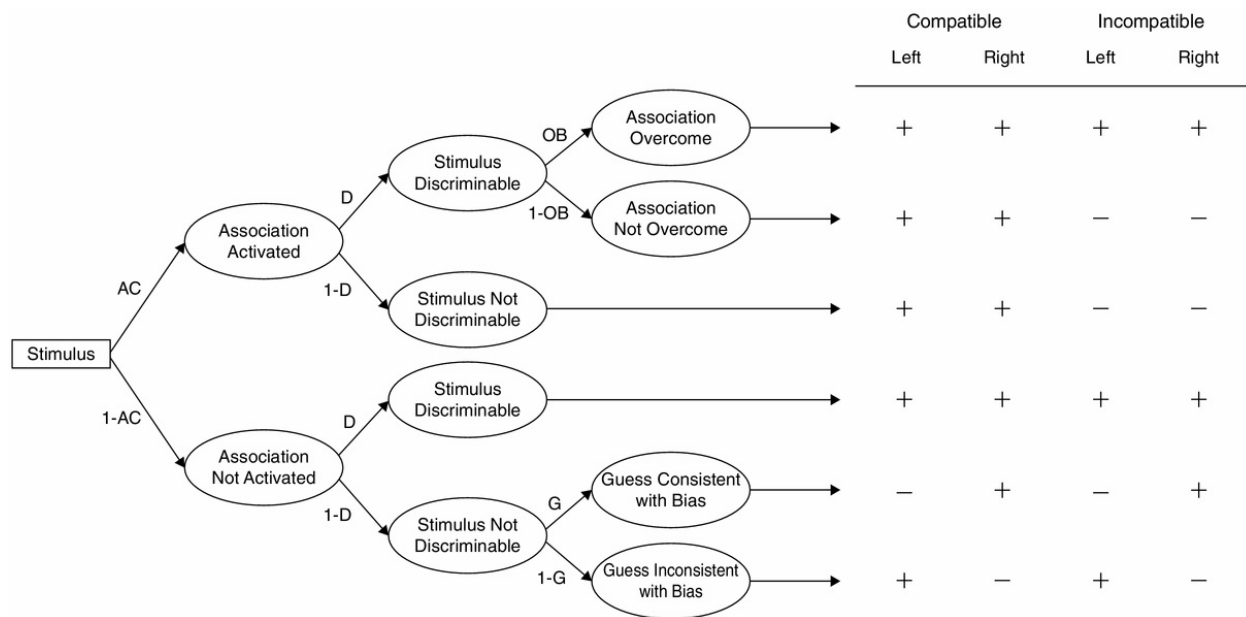


Figure 12.3. The quad-model of processes underlying correct (+) and incorrect (-) responses on indirect measurement procedures that are based on response interference. Figure adapted from Conrey, Sherman, Gawronski, Hugenberg, and Groom (2005). Reprinted with permission.

The contribution of multiple processes to responses on indirect measurement procedures has important implications for the interpretation of empirical findings. First, when using traditional measurement scores as independent variables (e.g., in studies on the prediction of behavior), the obtained relations to a criterion measure could be driven by an overlap in construct-unrelated processes. A potential example might be the correlation between an implicit measure of food attitudes and impulsive eating behavior, which could be driven by individual differences in the ability to inhibit unwanted response tendencies instead of genuine differences in food attitudes. Second, when using traditional measurement scores as dependent variables (e.g., in studies on attitude change), the measurement scores may be influenced by experimentally induced changes in construct-unrelated processes. For example, increased levels of prejudice on the race IAT after alcohol consumption have been shown to be the result of impaired inhibitory control rather than genuine changes in prejudice levels (Sherman, Gawronski, Gonsalkorale, Hugenberg, Allen, & Groom, 2008). Such ambiguities can be resolved by means of mathematical modeling techniques, such as the quad-model (Conrey et al., 2005) and other kinds of modeling procedures (e.g., Klauer et al., 2007; Payne, 2008).

Where Are We Going?

Up to now, method-focused research on implicit measures has primarily focused on the development of new measurement procedures and attempts to improve existing paradigms (Payne & Gawronski, 2010). For the decade to come, we believe that the field would benefit from a stronger focus on underlying mechanisms with regard to the measures themselves as well as their capability to predict behavior (see also Nosek, Hawkins, & Frazier, 2011). The groundwork for this focus has already been set by the development of mathematical modeling techniques (e.g., Conrey et al., 2005; Klauer et al., 2007; Payne, 2008), in which measurement outcomes are treated as behaviors that are themselves in need of a psychological explanation rather than as direct reflections of mental constructs (e.g., automatic associations) that can be used to explain behavior. As we outline in the remaining sections of this chapter, this perspective has several important implications.

Mechanisms Underlying Behavior Prediction

If the outcomes of psychological measurements are treated as behaviors rather than as direct reflections of mental constructs, one could argue that direct and indirect measurement procedures differ with regard to the processing constraints that they impose during the measurement of behavior. For example, traditional self-report measures of attitudes ask participants to intentionally evaluate the relevant attitude object, and the time for this evaluation is typically unlimited. In contrast, there is no requirement to intentionally evaluate the primes in an evaluative priming task, and participants are asked to respond as quickly as possible. Yet, when the similarity between the processing constraints of direct and indirect measures is increased (e.g., by imposing a time limit in the self-report measure), the correspondence of their measurement outcomes increases accordingly (e.g., Ranganath, Smith, & Nosek, 2008).

This idea can also be applied to the assessment of behavior. Specifically, one could argue that the predictive validity of implicit and explicit measures of the same construct should depend on the match versus mismatch of the processing constraints that are imposed by the measurement procedure and the processing constraints in the assessment of the to-be-predicted behavior (Fazio, 2007). Importantly, because indirect measurement procedures may differ with regard to the processing constraints in a given task, the same idea applies to the prediction of behavior by means of implicit measures. For example, when an indirect measurement procedure captures responses that are unintentional yet resource-

dependent, these responses might be a better predictor of behavior that is unintentional and resource-dependent. The same responses may be less suitable to predict behavior that is intentional but resource-independent.

Another implication is that predictive relations between psychological measures and observed behavior do not reflect the causal impact of a directly measured mental construct (e.g., automatic association) on the observed behavior, but rather covariations between two instances of behavior that are presumably driven by the same combination of processes and representations. Hence, successful prediction of behavior depends not only on the correspondence of the processing constraints in the measurement procedure and the to-be-predicted behavior, but on the entire set of processes that are involved in the production of the relevant responses. From this perspective, prediction of behavior by means of implicit measures might be improved by considering the conglomerate of processes that influence responses on the measurement procedure as well as the conglomerate of processes that underlie the to-be-predicted behavior. To the extent that indirect measurement procedures can be designed to match the combination of processes that are relevant in real-life situations, behavior prediction by means of implicit measures should be significantly improved.

To illustrate these arguments, consider the four processes proposed by Conrey et al.'s (2005) quad-model: the activation of an association-related response tendency (*AC*), the discrimination of the target stimulus (*D*), the success at overcoming association-related response tendencies in favor of the correct response (*OB*), and the impact of a general response bias (*G*). As we outlined earlier, all of these processes play a significant role in the IAT (and other measurement procedures based on responses interference; Gawronski et al., 2008). Although this lack of process purity may be regarded as a methodological flaw because of the implied confounds, it might in fact be functional for the prediction of behavior that is driven by the same combination of processes. For example, when a police officer has to make a split-second decision whether or not to shoot at a black suspect holding either a gun or a harmless object (Correll, Park, Judd, Wittenbrink, Sadler, & Keese, 2007), the officer's decision may be influenced by race-related associations between black people and guns (*AC*), the officer's ability to identify the object held by the suspect (*D*), the officer's success at overcoming an association-related tendency to pull the trigger (*OB*), and a general response tendency to shoot or not to shoot (*G*). Thus, to the extent that performance on an indirect measurement procedure involves all of these processes, its success in predicting decisions to shoot may be higher than when it

involves only a subset. Moreover, because the involved processes may be influenced by different situational affordances, the processing constraints in the indirect measure should be designed to match the ones in the to-be-predicted behavior. For example, the discriminability of the object held by the suspect may depend on visual conditions (e.g., daytime vs. nighttime), whereas success at overcoming an association-related tendency to pull the trigger may be reduced under time pressure. Ideally, both processing constraints should be equivalent in the measurement procedure and the to-be predicted behavior. The bottom line is that any behavioral response is the product of multiple different processes, and this idea applies to both responses on indirect measurement procedures and to-be-predicted behaviors. Hence, the predictive validity of indirect measures should be higher if their underlying processes and processing constraints match those of the to-be-predicted behavior.

Convergence versus Divergence between Implicit Measures

These considerations may also help clarify why different kinds of implicit measures sometimes show diverging effects. For example, a number of studies showed different effects of the same experimental manipulation on Fazio et al.'s (1986) evaluative priming task and Payne et al.'s (2005) AMP (e.g., Deutsch & Gawronski, 2009; Gawronski, Cunningham, LeBel, & Deutsch, 2010). From a traditional measurement perspective, these findings might be attributed to the mechanisms underlying different kinds of priming tasks, and these mechanisms may be distinguished from the to-be-measured psychological construct (e.g., automatic associations influence measurement outcomes by means of different task-specific mechanisms; Gawronski et al., 2008). However, if the outcomes of indirect measurement procedures are treated as behavioral responses, the mechanisms underlying a given measurement procedure become essential for understanding the production of the behavioral responses themselves.

To illustrate this argument, consider the task demands in Fazio et al.'s (1986) evaluative priming task and Payne et al.'s (2005) AMP. In the evaluative priming task, participants have to identify the correct response to the target stimulus, and the execution of this response might be facilitated or impaired by a valence-related response tendency elicited by the preceding prime (e.g., a response tendency to press the negative key elicited by a negative prime stimulus). From this perspective, priming effects are attributable to synergistic versus antagonistic effects of the response tendencies that are elicited by the primes and

the targets (Gawronski et al., 2008). This situation is quite different in the AMP, in which participants have to disambiguate the evaluative connotation of a neutral target stimulus. There is no correct or incorrect response in the AMP. In other words, whereas the evaluative priming task involves a situation of response conflict, the AMP involves a situation of evaluative disambiguation.

These considerations have important implications for the relation between the two tasks and their capacity in predicting behavior. For example, whereas the evaluative priming task might be a better predictor of behavior that involves the resolution of response conflicts (e.g., inhibition of an association-related tendency to pull the trigger of a gun in response to a black man holding an object that is identified as harmless), the AMP might be a better predictor of behavior that involves evaluative disambiguation (e.g., tendency to pull the trigger of a gun in response to a black man holding an ambiguous object). Moreover, the respective processes involved in the two kinds of responses may be differentially affected by the same factor, thereby leading to different outcomes of the same experimental manipulation. For example, attention to particular features of an attitude object may eliminate response conflicts resulting from evaluative connotations of irrelevant stimulus features. However, attention to particular features of an attitude object may be less effective in reducing the impact of irrelevant stimulus features on the processes that are involved in evaluative disambiguation. Consistent with these assumptions, Gawronski *et al.* (2010) found that attention to the category membership of face primes (i.e., age vs. race) moderated priming effects in Fazio et al.'s (1986) evaluative priming task but not in Payne et al.'s (2005) AMP.

The bottom-line is that responses on indirect measurement procedures are driven by different underlying mechanisms, and these mechanisms play an essential role in the production of the behavioral responses that are assessed by these procedures. Thus, to the extent that the involved mechanisms respond differently to the same situational influence, different measurement procedures may show diverging outcomes even when they are designed to assess the same psychological construct. Moreover, behavior prediction should be enhanced to the extent that the mechanisms underlying a given measurement procedure match the mechanisms underlying the to-be-predicted behavior .

Final Remarks

The validity of self-report measures is often challenged when people are

unwilling or unable to provide accurate reports of their own psychological attributes. This concern has been a driving force in the development of indirect measurement procedures. However, the evidence that has been gathered so far suggests a more complex relation between the two types of measures. Although social desirability and introspective limits may play a role for dissociations between explicit and implicit measures, researchers should be careful to avoid the fallacy of reverse inference by interpreting any dissociation in these terms (cf. Gawronski & Bodenhausen, in press). To avoid premature conclusions, we recommend that theoretical interpretations of measurement dissociations should be supported with relevant empirical data. Such data will not only provide deeper insights into why implicit and explicit measures show different antecedents and correlates; they may also advance the development of new measurement procedures and ultimately the prediction of behavior.

References

- Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, 84, 888–918.
- Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between explicit and implicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology*, 83, 380–393.
- Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, 7, 136–141.
- Banse, R., Gawronski, B., Rebetez, C., Gutt, H., & Morton, J. B. (2010). The development of spontaneous gender stereotyping in childhood: Relations to stereotype knowledge and stereotype flexibility. *Developmental Science*, 13, 298–306.
- Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the Affective Misattribution Procedure. *Personality and Social Psychology Bulletin*, 38, 1194–1208.
- Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The sorting paired features task: A measure of association strengths. *Experimental Psychology*, 56, 329–343.
- Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention,

- efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 1–40). Hillsdale, NJ: Erlbaum.
- Bargh, J. A., Chaiken, S., Raymond, P., & Hymes, C. (1996). The automatic evaluation effect: Unconditional automatic activation with a pronunciation task. *Journal of Personality and Social Psychology*, 32, 104–128.
- Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record*, 60, 527–542.
- Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, 61, 27–41.
- Bluemke, M., & Frieze, M. (2006). Do irrelevant features of stimuli influence IAT effects? *Journal of Experimental Social Psychology*, 42, 163–176.
- Brendl, C. M., Markman, A. B., & Messner, C. (2005). Indirectly measuring evaluations of several attitude objects in relation to a neutral reference point. *Journal of Experimental Social Psychology*, 41, 346–368.
- Briñol, P., Petty, R. E., & Wheeler, S. C. (2006). Discrepancies between explicit and implicit self-concepts: Consequences for information processing. *Journal of Personality and Social Psychology*, 91, 154–170.
- Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin*, 25, 215–224.
- Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. (2005). Separating multiple processes in implicit social cognition: The Quad-Model of implicit task performance. *Journal of Personality and Social Psychology*, 89, 469–487.
- Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, 92, 1006–1023.
- Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measurement: Consistency, stability, and convergent validity. *Psychological Science*, 12, 163–170.

- De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental Psychology*, 50, 77–85.
- De Houwer, J., & De Bruycker, E. (2007). The identification-EAST as a valid measure of implicit attitudes toward alcohol-related stimuli. *Journal of Behavior Therapy and Experimental Psychiatry*, 38, 133–143.
- De Houwer, J., & Moors, A. (2010). Implicit measures: Similarities and differences. In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 176–193). New York: Guilford Press.
- De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, 135, 347–368.
- Deutsch, R., & Gawronski, B. (2009). When the method makes a difference: Antagonistic effects on “automatic evaluations” as a function of task characteristics of the measure. *Journal of Experimental Social Psychology*, 45, 101–114.
- Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, 23, 316–326.
- Eder, A. B., & Rothermund, K. (2008). When do motor behaviors (mis)match affective stimuli? An evaluative coding view of approach and avoidance reactions. *Journal of Experimental Psychology: General*, 137, 262–281.
- Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, 25, 603–637.
- Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, 69, 1013–1027.
- Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, 50, 229–238.
- Field, M., Caren, R., Fernie, G., & De Houwer, J. (2011). Alcohol approach tendencies in heavy drinkers: Comparison of effects in a Relevant Stimulus-Response Compatibility Task and an approach / avoidance Simon task.

Psychology of Addictive Behaviors, 25, 697–701.

- Friese, M., Hofmann, W., & Schmitt, M. (2008). When and why do implicit measures predict behaviour? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology*, 19, 285–338.
- Gaertner, S. L., & McLaughlin, J. P. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly*, 46, 23–30.
- Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision makers. *Science*, 321, 1100–1102.
- Galdi, S., Gawronski, B., Arcuri, L., & Friese, M. (2012). Selective exposure in decided and undecided individuals: Differential relations to automatic associations and conscious beliefs. *Personality and Social Psychology Bulletin*, 38, 559–569.
- Gawronski, B. (2009). Ten frequently asked questions about implicit measures and their frequently supposed, but not entirely correct answers. *Canadian Psychology*, 50, 141–150.
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, 132, 692–731.
- Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology*, 44, 59–127.
- Gawronski, B., & Bodenhausen, G. V. (in press). Theory evaluation. In B. Gawronski, & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology*. New York: Guilford Press.
- Gawronski, B., Cunningham, W. A., LeBel, E. P., & Deutsch, R. (2010). Attentional influences on affective priming: Does categorization influence spontaneous evaluations of multiply categorizable objects? *Cognition and Emotion*, 24, 1008–1025.
- Gawronski, B., Deutsch, R., & Banse, R. (2011). Response interference tasks as indirect measures of automatic associations. In K. C. Klauer, A. Voss, & C.

- Stahl (Eds.), *Cognitive methods in social psychology* (pp. 78–123). New York: Guilford Press.
- Gawronski, B., Deutsch, R., LeBel, E. P., & Peters, K. R. (2008). Response interference as a mechanism underlying implicit measures: Some traps and gaps in the assessment of mental associations with experimental paradigms. *European Journal of Psychological Assessment*, 24, 218–225.
- Gawronski, B., Ehrenberg, K., Banse, R., Zukova, J., & Klauer, K. C. (2003). It's in the mind of the beholder: The impact of stereotypic associations on category-based and individuating impression formation. *Journal of Experimental Social Psychology*, 39, 16–30.
- Gawronski, B., Geschke, D., & Banse, R. (2003). Implicit bias in impression formation: Associations influence the construal of individuating information. *European Journal of Social Psychology*, 33, 573–589.
- Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are “implicit” attitudes unconscious? *Consciousness and Cognition*, 15, 485–499.
- Gawronski, B., & LeBel, E. P. (2008). Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology*, 44, 1355–1361.
- Gawronski, B., & Payne, B. K. (Eds.). (2010). *Handbook of implicit social cognition: Measurement, theory, and applications*. New York: Guilford Press.
- Gawronski, B., & Sritharan, R. (2010). Formation, change, and contextualization of mental associations: Determinants and principles of variations in implicit measures. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 216–240). New York: Guilford Press.
- Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology*, 40, 535–542.
- Gawronski, B., & Ye, Y. (2011). What drives priming effects in the affect misattribution procedure? Underlying mechanisms and new applications. Manuscript submitted for publication.
- Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and*

Social Psychology, 70, 491–512.

- Gray, N. S., Brown, A. S., MacCulloch, M. J., Smith, J., & Snowden, R. J. (2005). An implicit test of the associations between children and sex in pedophiles. *Journal of Abnormal Psychology*, 114, 304–308.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102, 4–27.
- Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109, 3–25.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85, 197–216.
- Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, 90, 1–20.
- Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, 135, 555–588.
- Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measure. *Personality and Social Psychology Bulletin*, 31, 1369–1385.
- Hofmann, W., Gschwendner, T., Nosek, B. A., & Schmitt, M. (2005). What moderates implicit-explicit consistency? *European Review of Social Psychology*, 16, 335–390.
- Hofmann, W., Rauch, W., & Gawronski, B. (2007). And deplete us not into temptation: Automatic attitudes, dietary restraint, and self-regulatory

- resources as determinants of eating behavior. *Journal of Experimental Social Psychology*, 43, 497–504.
- Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science*, 14, 640–643.
- Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The dominance of associative theorising in implicit attitude research: Propositional and behavioral alternatives. *The Psychological Record*, 61, 465–498.
- Jordan, C. H., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E., & Correll, J. (2003). Secure and defensive high self-esteem. *Journal of Personality and Social Psychology*, 85, 969–978.
- Karpinski, A., & Steinman, R. B. (2006). The Single Category Implicit Association Test as a measure of implicit social cognition. *Journal of Personality and Social Psychology*, 91, 16–32.
- Klauer, K. C., & Teige-Mocigemba, S. (2007). Controllability and resource dependence in automatic evaluation. *Journal of Experimental Social Psychology*, 43, 648–655.
- Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process-components of the Implicit Association Test: A diffusion model analysis. *Journal of Personality and Social Psychology*, 93, 353–368.
- Krieglmeyer, R., & Deutsch, R. (2010). Comparing measures of approach-avoidance behavior: The manikin task vs. two versions of the joystick task. *Cognition and Emotion*, 24, 810–828.
- LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, 37, 570–583.
- McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–126). New York: Academic Press.
- Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General*, 132, 455–469.

- Moors, A., & De Houwer, J. (2006). Automaticity: A conceptual and theoretical analysis. *Psychological Bulletin*, 132, 297–326.
- Nicholson, E., & Barnes-Holmes, D. (2012). The Implicit Relational Assessment Procedure (IRAP) as a measure of spider fear. *The Psychological Record*, 62, 263–278.
- Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition*, 19, 625–666.
- Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin*, 31, 166–180.
- Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences*, 15, 152–159.
- Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, 32, 421–433.
- Payne, B. K. (2008). What mistakes disclose: A process dissociation approach to automatic and controlled processes in social psychology. *Social and Personality Psychology Compass*, 2, 1073–1092.
- Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K. B. (2013). Intention invention and the Affect Misattribution Procedure: Reply to Bar-Anan and Nosek (2012). *Personality and Social Psychology Bulletin*, 39, 375–386.
- Payne, B. K., Burkley, M., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology*, 94, 16–31.
- Payne, B. K., Cheng, S. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89, 277–293.
- Payne, B. K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going? In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 1–15). New York: Guilford Press.

- Payne, B. K., Krosnick, J. A., PASEK, J., Lelkes, Y., Akhtar, O., & Tompson, T. (2010). Implicit and explicit prejudice in the 2008 American presidential election. *Journal of Experimental Social Psychology*, 46, 367–374.
- Penke, L., Eichstaedt, J., & Asendorpf, J. B. (2006). Single Attribute Implicit Association Tests (SA-IAT) for the assessment of unipolar constructs: The case of sociosexuality. *Experimental Psychology*, 53, 283–291.
- Perugini, M., Richetin, J., & Zogmaister, C. (2010). Prediction of behavior. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 255–277). New York: Guilford Press.
- Peters, K. R., & Gawronski, B. (2011). Mutual influences between the implicit and explicit self-concepts: The role of memory activation and motivated reasoning. *Journal of Experimental Social Psychology*, 47, 436–442.
- Petty, R. E., Briñol, P., & DeMarree, K. G. (2007). The meta-cognitive model (MCM) of attitudes: Implications for attitude measurement, change, and strength. *Social Cognition*, 25, 657–686.
- Petty, R. E., Fazio, R. H., & Briñol, P. (2009). The new implicit measures: An overview. In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 3–18). New York: Psychology Press.
- Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology*, 44, 386–396.
- Remue, J., De Houwer, J., Barnes-Holmes, D., Vanderhasselt, M.-A., & De Raedt, R. (in press). Self-esteem revisited: Performance on the implicit relational assessment procedure as a measure of self-versus ideal self-related cognitions in dysphoria. *Cognition & Emotion*.
- Richetin, J., Perugini, M., Adjali, I., & Hurling, R. (2007). The moderator role of intuitive versus deliberative decision making for the predictive validity of implicit measures. *European Journal of Personality*, 21, 529–546.
- Roddy, S., Stewart, I., & Barnes-Holmes, D. (2011). Facial reactions reveal that slim is good but fat is not bad: Implicit and explicit measures of body-size bias. *European Journal of Social Psychology*, 41, 688–694.

- Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Minimizing the influence of recoding in the IAT: The Recoding-Free Implicit Association Test (IAT-RF). *Quarterly Journal of Experimental Psychology*, 62, 84–98.
- Rudman, L. A. (2004). Sources of implicit attitudes. *Current Directions in Psychological Science*, 13, 79–82.
- Rudman, L. A., Dohn, M. C., & Fairchild, K. (2007). Implicit self-esteem compensation: Automatic threat defense. *Journal of Personality and Social Psychology*, 93, 798–813.
- Rydell, R. J., & Gawronski, B. (2009). I like you, I like you not: Understanding the formation of context dependent automatic attitudes. *Cognition and Emotion*, 23, 1118–1152.
- Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, 91, 995–1008.
- Rydell, R. J., McConnell, A. R., & Mackie, D. M. (2008). Consequences of discrepant explicit and implicit attitudes: Cognitive dissonance and increased information processing. *Journal of Experimental Social Psychology*, 44, 1526–1532.
- Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology*, 39, 590–598.
- Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately assess autobiographical events. *Psychological Science*, 19, 772–780.
- Scarabis, M., Florack, A., & Gosejohann, S. (2006). When consumers follow their feelings: The impact of affective or cognitive focus on the basis of consumer choice. *Psychology & Marketing*, 23, 1015–1034.
- Scherer, L. D., & Lambert A. J. (2009). Contrast effects in priming paradigms: Implications for theory and research on implicit attitudes. *Journal of Personality and Social Psychology*, 97, 383–403.
- Schnabel, K., Banse, R., & Asendorpf, J. B. (2006). Employing automatic approach and avoidance tendencies for the assessment of implicit personality

- self-concept: The Implicit Association Procedure (IAP). *Experimental Psychology*, 53, 69–76.
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54, 93–105.
- Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition*, 25, 638–656.
- Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. A., & Groom, C. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review*, 115, 314–335.
- Sherman, J. W., Klauer, K. C., & Allen, T. J. (2010). Mathematical modeling of implicit social cognition: The machine in the ghost. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 156–175). New York: Guilford Press.
- Solarz, A. K. (1960). Latency of instrumental responses as a function of compatibility with the meaning of eliciting verbal signs. *Journal of Experimental Psychology*, 59, 239–245.
- Spruyt, A., Hermans, D., De Houwer, J., Vandekerckhove, J., & Eelen, P. (2007). On the predictive validity of indirect attitude measures: Prediction of consumer choice behavior on the basis of affective priming in the picture-picture naming task. *Journal of Experimental Social Psychology*, 43, 599–610.
- Sriram, N., & Greenwald, A. G. (2009). The Brief Implicit Association Test. *Experimental Psychology*, 56, 283–294.
- Steffens, M. C. (2004). Is the Implicit Association Test immune to faking? *Experimental Psychology*, 51, 165–179.
- Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, 8, 220–247.
- Teige, S., Schnabel, K., Banse, R., & Asendorpf, J. B. (2004). Assessment of multiple implicit self-concept dimensions using the Extrinsic Affective Simon Task. *European Journal of Personality*, 18, 495–520.
- Teige-Mocigemba, S., Klauer, K. C., & Rothermund, K. (2008). Minimizing method-specific variance in the IAT: The Single Block IAT. *European*

Journal of Psychological Assessment, 24, 237–245.

- Teige-Mocigemba, S., Klauer, K. C., & Sherman, J. W. (2010). A practical guide to the Implicit Association Test and related tasks. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 117–139). New York: Guilford Press.
- Wentura, D., & Degner, J. (2010). A practical guide to sequential priming and related tasks. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 95–116). New York: Guilford Press.
- Whitfield, M., & Jordan, C. H. (2009). Mutual influences of explicit and implicit attitudes. *Journal of Experimental Social Psychology*, 45, 748–759.
- Williams, B. J., & Kaufmann, L. M. (2012). Reliability of the go/no-go association task. *Journal of Experimental Social Psychology*, 48, 879–891.
- Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, 107, 101–126.
- Wittenbrink, B. (2007). Measuring attitudes through priming. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 17–58). New York: Guilford Press.
- Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationships with questionnaire measures. *Journal of Personality and Social Psychology*, 72, 262–274.
- Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology*, 81, 815–827.

¹ For “cook-book” style instructions that include procedural information regarding the implementation of different paradigms (e.g., number of trials, presentation times, etc.), we recommend the chapters by Gawronski, Deutsch, and Banse (2011), Teige-Mocigemba, Klauer, and Sherman (2010), and Wentura and Degner (2010).

² Note that reliability estimates tend to be lower for second and subsequent IATs if more than one IAT is administered in the same session (Gawronski et al., 2011).

³ There is still no consensus about how estimates of internal consistency should be calculated for implicit measures (cf. Williams & Kaufmann, 2012). We recommend to split all critical trials of the task into two test-halves (e.g., first versus second half of all trials of an evaluative priming task) and to calculate two separate measurement scores on the basis of the two test-blocks (e.g., one priming score on the basis of the first half and another one on the basis of the second half). The two scores can then be used to calculate a split-half coefficient or a Cronbach's Alpha value. Note that it is not appropriate to calculate reliability estimates on the basis of the raw data from different types of trials (e.g., mean responses latencies on different kinds of prime-target combinations). Such estimates would reflect the internal consistency of responses on different types of trials (e.g., internal consistency of responses latencies for positive and negative words), not the internal consistency of the implicit measurement score (e.g., internal consistency of evaluative priming effect).

⁴ For more information about the use of mathematical modeling techniques in research using implicit measures, we recommend the introductory overview by Sherman, Klauer, and Allen (2010). For more detailed information about particular modeling procedures, readers may consult Conrey *et al.* (2005), Klauer *et al.* (2007), and Payne (2008).

⁵ Note that the quad-model has been designed for indirect measurement procedures that are based on response interference. It is not applicable to tasks that are based on other mechanisms, such as the AMP (Gawronski, Deutsch, LeBel, & Peters, 2008).

Chapter thirteen The Mind in the Middle

A Practical Guide to Priming and Automaticity Research

John A. Bargh and Tanya L. Chartrand*

For most of the 20th century, it was not permissible to invoke cognitive mechanisms and processes to explain and model psychological phenomena. The difficulty was mainly methodological: A century ago, the only known research technique to examine these internal mental states was introspection and self-report. Introspection fell into disfavor because an objective outside observer could not independently verify the data by making the same measurements. Other sciences did not confuse the observer with the observed, it was said, and so neither should psychology if it wanted to be considered a science as well.

The dominant Behaviorist research paradigm initially eschewed the study of consciousness because of this lack of objective methods; however, over time this position became more extreme – that which could not be studied became that which did not exist at all (Bargh & Ferguson, 2000). And so the study of internal cognitive processes as mediators and moderators of the influence of the environment on higher mental processes, such as those involved in social interactions, was studiously neglected until the 1970s. In his book on behaviorism and the cognitive revolution, *The Ghost in the Machine*, Arthur Koestler (1967) observed that this dedicated disinterest in its natural subject matter – the internal life of the mind – caused psychology to go nowhere for decades, at a time in history when the other sciences, in contrast, were making giant strides forward.

Things are different today, of course. The research methods and techniques described in this chapter are a major reason why we now have a scientific social cognitive psychology. These methods do not rely on introspection and self-report; measurements are made by outside observers and are replicable by other outside observers. Instead of introspection, as a field we have learned how to make inferences about cognitive process and structure from response latencies, from the order in which our participants recall stimuli about people and events, and from what happens when the mental system is put under load or time pressure, as when the individual has to do several different things at once. For

the most part, we no longer have to rely solely on the person's own description of their internal state. Like the nuclear physicist inferring atomic structure from lines on a photographic plate, we can infer mental structure from 25-millisecond differences in the time taken, for example, to pronounce a specific word.

One cannot “see” inside another's mind, but neither can the physicist “see” quarks and muons inside the atom. Just as do other sciences, we infer, deduce, and build theories about the mind based on observables, generating falsifiable predictions and putting them to the test. (And we can even use introspection and self-report because we are able to verify and check these data against the other, independent data sources.)

The present chapter is a summary of the methods commonly used to explore the cognitive representations and processes that mediate between environmental events and human reactions to them – be those responses impressions, judgments, evaluations, emotions, goals, or behavior. We focus primarily on passive, or unintentional, forms of cognitive mediation in an attempt to keep it distinct from motivational mediation as much as possible. Goal effects on information processing and behavior are purposive and strategic (by definition) and not strictly attributable to cognitive structure or process per se (for a comprehensive review, see Gollwitzer & Moskowitz, 1996). But motivation and cognition are highly, if not inextricably, related (Gollwitzer & Bargh, 1996; Sorrentino & Higgins, 1986), and the reader of the present chapter will find many references to the intersection of the two – for instance, in the unintended carryover effects of a goal chosen intentionally in one context to a subsequent context. By maintaining a focus on passive or unintentional effects, we hope to keep to the theme of how to study the mental representations and processes that mediate and moderate social psychological phenomena.

Priming and automaticity research techniques share a concern with the ways that internal mental states mediate, in a passive and hidden manner, the effects of the social environment on psychological and behavioral responses. Automaticity techniques enable an experimenter to measure the particular mental procedures or representations that are assumed, in his or her theory, to correspond to the individual differences in a phenomenon. For example, Dodge (1993) has argued that violent boys differ from other boys in the ways that they automatically perceive the aggressive intentions of others. Many depression researchers, starting with Aaron Beck (1967), have proposed that depressed individuals tend to automatically think of themselves in negative terms and so suffer low self-worth, without having much awareness of how those feelings come about.

Priming studies, on the other hand, are more concerned with effects of the current situational context and how these environmental features cause the average individual to think, feel, and behave differently than otherwise.

Today, nearly four decades after Mischel (1973) proposed the mergence of social and personality psychology – that is, the study of individual differences in reactions to situational forces – the existence of individual differences in perception is well established in the field. Yet less than 70 years ago it was a radical thing (in experimental psychology) to suggest that one's experience of the outside world was determined by anything other than the stimulation “out there.” We start our treatment of cognitive research methods by presenting a brief history of cognitive mediation in psychology: first, the breakthrough idea that people could differ in what they perceived in the environment, and how they perceived it, followed by the various reasons found for these individual variations. The mind was not always in the middle of psychological explanation; here is how it got there.

The Influence of Internal States on Perceptual Experience

The early elementalist approach of Wundt and Titchener held that perception was explicable entirely in terms of discrete sensory events; indeed, any reference to perception of objects per se by the introspecting perceiver (instead of to the sensory features present in that object) was held to be going beyond the information present – an inference, not something actually perceived (Boring, 1950). The Gestalt movement, in fact, arose in direct opposition to the elementalist approach. The Gestaltists argued that people did indeed go beyond the information given, perceiving objects as wholes according to precise principles of form and relations that were not reducible to the sensory stimulation alone (Koffka, 1922).

The study of visual illusions provided the Gestalt movement with many powerful demonstrations that these emergent properties of the stimulus – and not merely the actual stimulus present – produced perceptions of size, distance, and brightness (Boring, 1950). For example, a black-and-white photograph of a woman in a white dress, standing next to a man wearing a dark suit, appears phenomenally the same under varying lighting conditions. This is despite the fact that the dark suit under the brighter lighting is actually the same (physically speaking) shade of grey as the white dress had been under the darker lighting.

Individual Differences in Perceptual Experience

When Christian Dior launched his “New Look” in fashion design in 1947, little did he know that he was also supplying the name for a radical movement in human perception research. What we know now as the “New Look” in perception was a break from the then-dominant assumption that perceptual experience was determined solely by properties of the stimulus field (including the Gestaltists’ emergent properties). For the first time, it was proposed that there could be individual differences in perceptual processing.

Although the Gestaltists showed that people go beyond the information present in the environment, the mechanisms by which they did so were still regarded as universal. Individual variation around the grand mean of judgments of intensity or other stimulus features was treated as error variance. But it had been noticed that there were consistent individual differences in these errors. Some experimental participants were consistently on the low side of the mean, with others usually on the high side, and this became known, somewhat oxymorically, as the constant error.

If these deviations had been merely random noise, a given individual would have been expected to vary randomly – not systematically – around the mean in his or her judgments. Recognizing this, Bruner and Postman (1947) proposed that these constant errors were not errors at all, but true individual differences in perceptual experience. Moreover, they surmised that the observed individual differences were perhaps correlated with other individual differences, such as in motivations, needs, and values. The New Look in perception was born.

Suddenly, entire areas of psychological inquiry – attitudes and values, emotion, motivation and goal research, personality, clinical and psychodynamic theory – had a bridge to experimental psychology. New Look research boomed as these researchers explored the effects of their particular brand of individual difference on perceptual experience (see reviews by Allport, 1955; Bruner, 1957; Dixon, 1971; Erdelyi, 1974). In a very real way, it was the birth not just of a fruitful avenue of perception research, but of a truly general experimental psychology, as the laboratory techniques that had long been associated with “scientific” psychology could now be exploited by these other areas.

The Roots of Priming Research

Although the New Look championed the role played by individual differences in

motives and needs in perceptual experience, nowhere in it was there a mention of what we now refer to as *priming* – how recent or current experience passively (without an intervening act of will) creates internal readinesses. Bruner's (1957) classic statement of category accessibility theory described how current goals and purposes caused representations relevant to achieving those goals to become more accessible and ready to be activated by their corresponding objects and events in the environment. But this was a quite active and intentional internal state.

Recent Experience as an Individual Difference

A closer historical precedent to present-day research on passive contextual effects is Duncker's (1945) pioneering work on mindsets and creativity. Duncker showed that a person's usual way of thinking about objects and their functions sometimes gets in the way of coming up with novel, creative solutions to problems. For example, imagine Joe is given the task of tying together two pieces of string dangling far enough apart that he cannot grasp the one piece without letting go of the other. Joe also has a hammer at his disposal, but on his own he cannot figure out how it could help him to complete the task. However, as soon as the experimenter sets one of the dangling threads into motion, it occurs to Joe to tie the hammer to the end of the string in order to set it in motion like a pendulum. Importantly, Joe is not aware of the effect that the experimenter's knocking the string into motion had on his (Joe's) arrival at the correct solution. Although not framed this way at the time, today we understand this phenomenon as a case of passive conceptual priming – the concept, in this case, being that of motion. This activated concept becomes more likely than before to influence conscious judgments and problem solving.¹

The first use of the term “priming” to refer to the temporary internal activation of response tendencies was by Karl Lashley in a 1951 article. Lashley was dealing with the problem of how serial response sequences, as in speech production, flow so quickly and apparently effortlessly. He argued that there had to be a mediating state intervening between the act of will or intention and the production of the intended behavior, which assembled the action into the proper serial sequence. This he called the priming of the response. The idea of priming thus entered the literature to refer to a preparedness of mental representations to serve a response function, although the activation Lashley described came from internal, and even intentional, sources.

In the first demonstration of passive priming within a memory paradigm,

Storms (1958) first gave his participants a list of words to memorize and then had them free-associate to a series of stimulus words. Unexpectedly, Storms found that the words presented in the memory task became more likely than usual to be given as associates (compared with standard free-association norms). Storms reported this effect but could not explain it, concluding that “the mechanisms of this recency effect remain unexplored” (p. 394).

It was Segal and Gofer (1960) who first used the term “priming” to refer to this effect of recent use of a concept in one task on its probability of usage in a subsequent, unrelated task. Segal and Gofer replicated Storms's finding but, critically, without the use of explicit recall instructions; merely exposing participants to the list of words had the effect of increasing the probability that those words would be used in the subsequent free-association task.

Following this initial demonstration, priming began to be used as an experimental technique, especially to show how information had been stored in memory despite the individual's inability to recall it (Grand & Segal, 1966; Koriati & Feuerstein, 1976; Segal, 1967). That is, words presented in a first task still were more likely than usual to show up as free associates in a subsequent task, even though participants had failed to recall them at the end of the first task. These early priming studies were thus the forerunners of the distinction between implicit and explicit forms and uses of memory (e.g., Greenwald & Banaji, 1995; Schacter, 1987).

Priming in Social Psychology

For social psychology, the groundbreaking priming study came when Higgins, Rholes, and Jones (1977) showed that personality trait concepts (such as “adventurous” or “independent”) – not just single words – could be primed by recent use. Using the same unrelated studies paradigm as had Segal and his colleagues, Higgins *et al.* (1977) exposed participants to synonyms of certain personality traits as part of an initial memory experiment. Next, in what participants believed to be an unrelated experiment, they read about a target person named Donald who behaved in ways ambiguously related to the primed traits, such as sailing across the ocean alone and preferring to study by himself. Those participants who had been exposed to words such as “adventurous” and “independent” formed more positive impressions of Donald than did participants who had been previously exposed to terms such as “reckless” and “aloof.” Importantly, participants evidenced no awareness of having been influenced by their prior exposure to trait terms in the earlier memory experiment.

An important advance beyond previous priming studies was that the participants' responses did not involve using the prime words themselves, as in the free-association task studies; instead their overall impression or evaluation of Donald was requested. What had been primed, therefore, were not just the single, concrete lexical memory locations corresponding to the stimulus words, but also the abstract trait concepts. These in turn became more likely to capture the relevant but ambiguous behavioral information, thus slanting final impressions in the positive or negative direction.

The Higgins *et al.* (1977) study revealed for the first time how an individual's recent experience could affect – in a passive and unintended way – his or her perceptual interpretation of another person's behavior. In their study, all participants read about the same target person doing the same things, yet they came away from their reading with markedly different impressions of that person, differences that could only be explained by the manipulated differences in their recent use of the trait concepts.

The Roots of Automaticity Research

Priming and automaticity research have a common purpose: to explore the effects of individual differences in accessibility of mental representations on perception, evaluation, motivation, and behavior. However, whereas priming research centers on the temporary activation of an individual's mental representations by the environment and the effect of this activation on various psychological phenomena, automaticity research focuses on more permanent, “hard-wired” sources of activation – that is, chronic accessibility of social knowledge structures. We now turn to the development of the present-day conception of automaticity.

It is now widely held that automatic processing is not a singular entity, but rather a grab-bag of the various types of processing that are considered *not conscious* (Bargh, 1989, 1994, 1996; Logan & Cowan, 1984; Neumann, 1984; Wegner & Bargh, 1998). That is, although there has been consensus over the years as to the qualities of deliberate or controlled processing, different kinds of “not-conscious” processes have been noted and studied. Conscious processing, by all accounts, is serial (sequential), rather than parallel, in nature; is limited in the amount of information it can handle at any one time; corresponds roughly to the contents of phenomenal awareness; and is directed by the individual's intentions and goals. The latter quality enables control processing to be flexible

and strategic and able to override (nearly always) the usual or habitual response in a situation.

And so, if a process or effect was discovered that did not have one or more of these features, it was considered to be “automatic” under the assumption that there were two and only two basic types of information processing: conscious and automatic (see, e.g., Johnson & Hasher, 1987; Posner & Snyder, 1975; Shiffrin & Schneider, 1977). Over the past century of research, however, two distinct strains of not-conscious processing have been discovered and studied. These two separate programs of research have led today to two major types of automaticity: goal-dependent and preconscious.

Goal-Dependent Automaticity and Skill Acquisition Research

One type of not-conscious processing concerns acquired skills that through a great deal of practice or experience come to be executed very efficiently, needing minimal if any attention or guidance (Newell & Rosenbloom, 1981; Shiffrin & Schneider, 1977). Examples of such skills are driving and typing, abilities that can operate without conscious guidance once started, but which are nonetheless intentional in that they require an act of conscious will to begin operation.

Although William James was not fond of the nonconscious as a scientific construct, his concept of habit did provide the heritage for modern-day conceptions of automaticity. James (1890) placed great importance on habit in daily life and believed that habits are ingrained by consistent and diligent practice. James's notion that activities frequently and consistently engaged in require less and less conscious effort over time became the foundation of skill acquisition research (Anderson, 1982; Newell & Rosenbloom, 1981).

For example, Shiffrin and Schneider (1977) proposed that perceptual skills can become automatized over time. They conducted a series of studies in which the participants' task was to detect a single letter or digit target within a rapidly presented array of letters and digits. After thousands of such trials, attention was automatically directed to the target. This pointed to the importance of frequency for the development of automaticity. Shiffrin and Schneider (1977) also showed the importance of consistency, in that automatic detection capabilities were only achieved when a stimulus was always a target or always a distracter; when the participants' response to the target varied, automatic responses did not develop.

It is important to note that in all of the skill acquisition research, past and present, there is an underlying assumption that an initial conscious act of will is required to set the effects into motion. One does not drive, type, or find targets in a perceptual display without having the intention of doing so, regardless of how efficient and automatic the processing is once one is engaged in the activity. This form of automaticity is called *goal-dependent* (Bargh, 1989) because, unlike the other major form (see the next section), it requires an initial intention or act of will to put the process into motion.

Preconscious Processing

The New Look was concerned with immediate reactions to a stimulus prior to it reaching conscious awareness. Today, the idea that a substantial amount of information processing occurs immediately on an environmental event – for instance, the activation of an individual's stereotype of a social group on the mere presence of a member of that group – has found wide acceptance. But at the time, the New Look's focus on motivational and personality determinants of conscious perceptual thresholds was very controversial. This was owing to its notion of perceptual defense, which, with its basis in Freudian notions of defense mechanisms, argued that perceptual thresholds were higher for emotionally threatening stimuli. However, if this were true, it would have violated the ingrained and implicit assumption of the time that perception was a conscious act (Erdelyi, 1974). The New Look's ideas about preconscious analysis were about 25 years ahead of their time, but eventually the assumption that all of perceptual activity is fully conscious was overthrown.

Mainly this occurred through research on selective attention, beginning with Broadbent's (1958) seminal work. Broadbent held that an individual is equipped with an internal, and intentionally operated, selection mechanism that “tunes” attention to focus on certain information in the environment and to disregard other information. But whereas Broadbent argued for an “early selection” theory of attention – that is, information to be selected is determined very early, prior to a complete analysis of the input for meaning – Treisman (1960) demonstrated that, in fact, some to-be-ignored contents do receive analysis for meaning, prior to attentional selection. Although her participants were very good at ignoring the to-be-unattended ear in a dichotic listening task, in which they were to repeat out loud a story played to one ear but not the other, there nonetheless were times when they would repeat the contents of the unattended channel. This occurred when the attended story switched to that ear. Thus the idea of pure early

selection was dispelled. Such a theory could not account for the switching of attention to an unattended channel based on the meaning of the information presented there.

After this demonstration that some selection outside of conscious awareness does indeed occur, a lively debate began as to exactly how much. Although many argued for a relatively early-selection model, in which only a limited amount of informational input is analyzed for meaning (e.g., Neisser, 1967), others (e.g., Deutsch & Deutsch, 1963) argued for a late-selection, complete analysis model. According to this model, there is a full and complete preconscious analysis of all sensed information for meaning and importance, and information enters consciousness depending on its importance for the individual.

In a historic synthesis of the two positions, Norman (1968) proposed that the extent of preconscious analysis varied, depending on the match between the external information and the readiness or accessibility of internal memory representations relevant to that information. And so cognitive psychology had come full circle back to the discredited ideas of the New Look, which had originally proposed that individual differences in internal states (i.e., attributable to emotions, needs, and goals) affected perception prior to the attainment of the conscious percept.

Finally, in a highly influential paper, Posner and Snyder (1975) put forth a model suggesting that automatic processes at encoding are triggered directly by the presence of the relevant stimulus. However, strategic conscious processes can override automatic ones to determine the response to the stimulus when the responses suggested by the two are incompatible – but only if the conscious process has enough time to develop and attentional capacity to operate. The basic proposals of this model were supported in a series of experimental tests by Neely (1977).

Priming and Automaticity Together

This is the heritage of contemporary priming and automaticity research in social and personality psychology. Priming studies are concerned with the temporary activation states of an individual's mental representations and how these internal readinesses interact with environmental information to produce perceptions, evaluations, and even goal pursuits and social behavior (see Bargh, 1997, 2006). Automaticity research is conceptually quite similar to priming studies, but generally concerns chronic individual differences in mental representations that

transcend the current context. Both types of research focus on the accessibility or ease of activation of social knowledge structures and how these influence psychological phenomena without the individual being aware of or intending such influences.

Moreover, because priming produces for a short time a level of activation and accessibility in a representation that is comparable to that of a long-term, automatic process (Bargh, Bond, Lombardi, & Tota, 1986), priming techniques also have been exploited as a way to experimentally manipulate what are theoretically posited as chronic, automatic effects. (For examples of this use of priming, see Bargh, Raymond, Pryor, & Strack, 1995, Experiment 2; Chen, Shechter, & Chaiken, 1996; Fazio, Sanbonmatsu, Powell, & Kardes, 1986, Experiment 3; Roskos-Ewoldsen & Fazio, 1992.) Thus priming techniques can be used either to research the passive, unintended influences of the current and recent environmental context or to experimentally simulate automaticity effects.

Priming Research Techniques

There are a variety of experimental techniques that fall under the general umbrella of priming research. *Conceptual priming* involves the activation of mental representations in one context, so that they exert a passive, unintended, and unaware influence in subsequent unrelated contexts until their activation dissipates. Examples of such research are the many trait concept priming studies in which, for instance, using the word “honest” as part of a language test causes one to later perceive a subsequent target person as being more honest. In this research, the participants’ task in processing the concept-relevant information (i.e., the priming task) is not the same – in fact is kept as different as possible – as their task in the subsequent part of the experiment that assesses the priming effect. In this way, the priming effect is shown to stem from the concepts primed (independent of processing goal) and not the priming of a particular mental procedure, which distinguishes this type of priming from the next.

Mindset priming manipulations have the participant actively engage (or read about someone else so engaged) in a goal-directed type of thought in one context, to show that this mindset (Gollwitzer, 1990) – what goal to pursue in the situation – is more likely to operate later in an unrelated context. Thus, what is primed is a procedure or way of thinking about information or a situation. For example, Wilson and Capitman (1982) had some of their male participants read a “boy meets girl” story in an allegedly unrelated first experimental task, and

they smiled more and generally behaved in a more friendly way to a female confederate in the next part of the experiment.

Unlike the other two types of priming studies, *sequential priming* techniques do not examine the residual effects of recent experience. Rather, they test for chronic connections between two representations, across which activation automatically spreads, for example between an attitude object and its evaluation or between two different concepts. Sequential priming is therefore the technique of choice for studying the associative structure of the mind. The discussion of sequential priming techniques therefore is postponed until the section on automaticity research, which is also concerned with long-term structural effects.

What all three types of priming have in common is a concern with the unintended consequences of an environmental event on subsequent thoughts, feelings, goals, and behavior. They address the residual effects of one's use of a representation in comprehending or acting on the world, which leaves the primed representation, or any other representation strongly (automatically) associated with it, active for some time thereafter. During the time it remains active, it exerts a passive effect on the individual, one that he or she is not aware of and does not intend – and is therefore unlikely to control (Bargh, 1994; Bruner, 1957; Higgins, 1989, 1996).

Conceptual Priming

In conceptual priming, manipulations are used that activate the internal mental representation of interest in a first task, in such a way that the participant does not realize the relation between that activation event and the later influence or use of that representation in an unrelated context. The priming task must use the concept or representation in some way, but not in a way that tips the participant off to the relation between the two tasks. To show it is just the mere activation of the representation that is necessary, and not its particular use in, say, person perception, tasks have commonly exposed the participant to representation-relevant stimuli (i.e., words or pictures) in an unobtrusive way.

Supraliminal Priming

There are different degrees to which an individual may be aware (or unaware) of the actual stimuli priming a given construct. In supraliminal priming, the participant is exposed to the priming stimuli as part of a conscious task. That is, the individual is fully aware of the priming stimuli itself but is not aware of

some underlying pattern that serves to prime the construct. A very frequently used supraliminal priming technique is the “scrambled sentence test,” first devised by Costin (1969) as a clinical projective test but adapted by Srull and Wyer (1979) in their trait construct priming research. Two examples are given in [Appendix A](#) at the end of the chapter. Participants are told that the task is designed to measure their language ability, and they are instructed to make coherent, grammatical sentences out of each string of words. In the course of doing so, they are exposed to some words or phrases that are related to the concept the experimenter wishes to prime.

When priming a relatively concrete concept (e.g., trait of honesty; goal to achieve), one can usually select priming stimuli by consulting a standard thesaurus for close synonyms of the to-be-primed concept. Pretesting can also be used to supplement this set of synonyms if more or varied priming stimuli are needed, by having a separate group of participants rate the degree to which each potential prime is related to the target concept. It is a good idea to use as many different words that are synonyms of the target concept in the scrambled sentence test as possible, because repeating a given word increases the chances that the participant may clue in to the purpose of the task, or at least become consciously aware that the experiment seems to be focusing on that particular concept.² At the same time, one must be careful not to sacrifice direct activation of the single concept of import by using only peripherally related primes.

Note that each prime word in this case can either be positioned as the word that is to be deleted (i.e., not used when creating the sentence) or can be one of the words used in forming the sentence. A mix is usually best, but only if the meaning of the sentences containing prime words (once unscrambled) does not give away the theme or the construct being activated in the sentences.

Sometimes the construct to be primed is not simple or concrete enough to simply use a list of synonyms. For instance, priming a goal to be entertained or to present oneself well, or priming a stereotype of nurses or skinheads, may be too abstract to rely solely on single words to activate the concept. In these cases, phrases can be used within the sentence to activate the concept. For example, Chartrand, Huber, Shiv, and Tanner (2008) used phrases such as “penny pincher,” “price conscious,” and “save money” to activate a thrift goal. Taking it one step further, the meaning of the sentence as a whole can be what is activating the concept (e.g., “He opposed the decision” and “Mary and Sally argued” to activate an antagonistic mindset). In this case, however, one has to be careful to include plenty of filler sentences in order to dilute the concept and

keep it below the participants' radar.

Generally speaking, one wants to have the most powerful manipulation possible, while at the same time not overstepping the line that leads to the participant's potential awareness of the experimental hypothesis. There is no cut-and-dried rule to achieve the "right" level of subtlety, but we can offer a few guidelines based on experience. One is to engage in extensive debriefing of the participant to ensure he or she is not cognizant of the relation between the priming manipulation and the subsequent experimental task. The best way of doing this is through a "funneled debriefing" (Chartrand & Bargh, 1996; Eagly & Chaiken, 1993). Appendix B gives an example of this technique. Briefly, the idea is to probe in a systematic way for any suspicions or actual knowledge the participant has about the intended effect of the prime on their subsequent performance in the experiment. The questions start out general and broad (e.g., "What did you think this study was about?") and get progressively more specific and focused on the priming manipulation (e.g., for a scrambled sentence task, "Did you notice any pattern or theme to the words in the scrambled sentences?").

In general, if a participant evidences any genuine awareness of a relation between the prime and experimental task, his or her data should not be included in the analyses. By "genuine awareness" we mean any answer in the debriefing that is "in the ballpark" as to what could have affected responses. In our research, we take a conservative stance and err on the side of overexclusion if there is any doubt.

If an alarmingly high proportion of participants are being excluded for this reason – and those alarms should go off if upward of 5% or so are showing awareness of the priming influence on their responses – it is likely that even participants who remain in the data set might have had some degree of awareness of the targeted content of the scrambled sentence test or even the experimental hypothesis.

The second tactic that we recommend is to replicate priming effects that are obtained with the more "conscious" or supraliminal priming techniques (e.g., the scrambled sentence test) using subliminal prime presentation instead (see next section). Although subliminal priming is a weaker manipulation, obtaining the same significant effect using it goes a long way toward dispelling doubts about the "demand" or conscious, strategic nature of the obtained priming effects. For example, in the Bargh *et al.* (1996) experiments on stereotype priming effects on behavior, Study 2 used the scrambled sentence test, but Study 3 used subliminal presentation of faces of the stereotyped-group in order to prime the stereotype,

with both methods producing stereotype-consistent priming effects on the participants' behavior.

The third tactic we strongly recommend is to ensure that the experimenter running the study remains blind to the experimental hypotheses and especially the priming condition of the participant in each session. For example, in the “elderly priming” experiment (Study 2) of Bargh *et al.* (1996), which used the scrambled sentence test to prime the elderly stereotype, for each session the test was placed inside an envelope and handed by the researcher to the session experimenter. He or she then gave this to the participant to complete privately in a room with the door closed, so that the experimenter was not aware of what the participant was filling out, and thus not aware of to which priming condition (elderly vs. control) the participant had been randomly assigned. Such care in keeping the session experimenter, who has contact with the study participants, blind to the study hypotheses and especially the participant's experimental condition is very important because it has long been known that experimenter's knowledge of hypotheses can sometimes produce the hypothesized effect in often quite subtle ways (Rosenthal, 1966). Indeed, Doyen, Klein, Pichon, and Cleeremans (2012) have recently claimed that such effects as the elderly stereotype priming effect on behavior stems entirely from the experimenter's awareness of the participant's priming condition, and the consequent subtle differences in how those participants were treated. However, their criticism cannot account for the effect observed in Bargh *et al.* (1996, Study 2) precisely because in those studies, as explained earlier, the appropriate steps were taken to ensure that the experimenter remained unaware of the participant's condition.

Subliminal Priming

Subliminal priming studies in social psychology can be useful, therefore, not only to demonstrate effects of nonconsciously perceived stimuli, but also to conclusively rule out alternative explanations for priming effects, such as experimental demand and participants becoming aware of the study hypotheses. (Discussion of subliminal priming is also relevant to the topic of automaticity, in its sense of processing without awareness; Bargh, 1994.) It was for this reason that Bargh and Pietromonaco (1982) performed the first subliminal trait construct priming study – to ensure that the original findings of Higgins *et al.* (1977) and Srull and Wyer (1979, 1980) had not been a result of demand or other active strategies on the part of the experimental participants. All of those previous studies had presented the critical primes to participants as part of a first,

explicit task. Similarly, Devine's (1989) use of the same procedure to prime the African-American stereotype was motivated by a wish to eliminate self-presentational strategies on the part of the experimental participants that could mask the true activation effects of the stereotype.

The mechanics of conducting subliminal priming studies are straightforward, and hinge on three principles: (1) very brief presentation of the prime, (2) its immediate masking by another stimulus, and (3) appropriate awareness checks.

Brevity of presentation translates into the amount of internal activation of the corresponding representation. Roughly speaking, the amount of internal activation is given by the formula $D \times I = A$, where D is the duration of the stimulus, I is its intensity, and A is the amount of activation. Using a tachistoscope, as was used in many of the New Look and perceptual microgenesis studies, one could vary the illumination level of the stimulus or use gelatin filters (similar to the effect of sunglasses) to make the stimulus harder to see. But the great majority of subliminal presentations in modern research accomplish their purpose by varying the duration, not the intensity, of the stimulus.

How long can a stimulus be presented and still be subliminal? Given that recognition thresholds are often, if not usually, measured in terms of millisecond duration and that there are individual differences in these thresholds (Greenwald, Klinger, & Liu, 1989), no single answer can be given. Establishing individual thresholds is a laborious and time-consuming process (e.g., requiring 30 minutes adaptation to the dark; Greenwald et al., 1989), so the practical solution is to use a duration brief enough to be outside the awareness of all or nearly all participants, and to conduct a conservative awareness check on these same participants (more on this later in the chapter).

The appropriate duration depends on whether the stimulus will be masked and whether it is presented to the participant's foveal or parafoveal visual field. Roughly speaking, *foveal processing* is given to information in the center or focus of conscious visual attention, and *parafoveal processing* applies to information on the fringe or periphery of the attended region. The foveal processing area extends from 0 to 2 degrees of visual angle from the focal point of attention (see Figure 13.1). In experiments involving a tachistoscope or a computer screen, foveal presentation is accomplished through presenting a "fixation point" (such as an asterisk) and the critical stimulus at the same position.

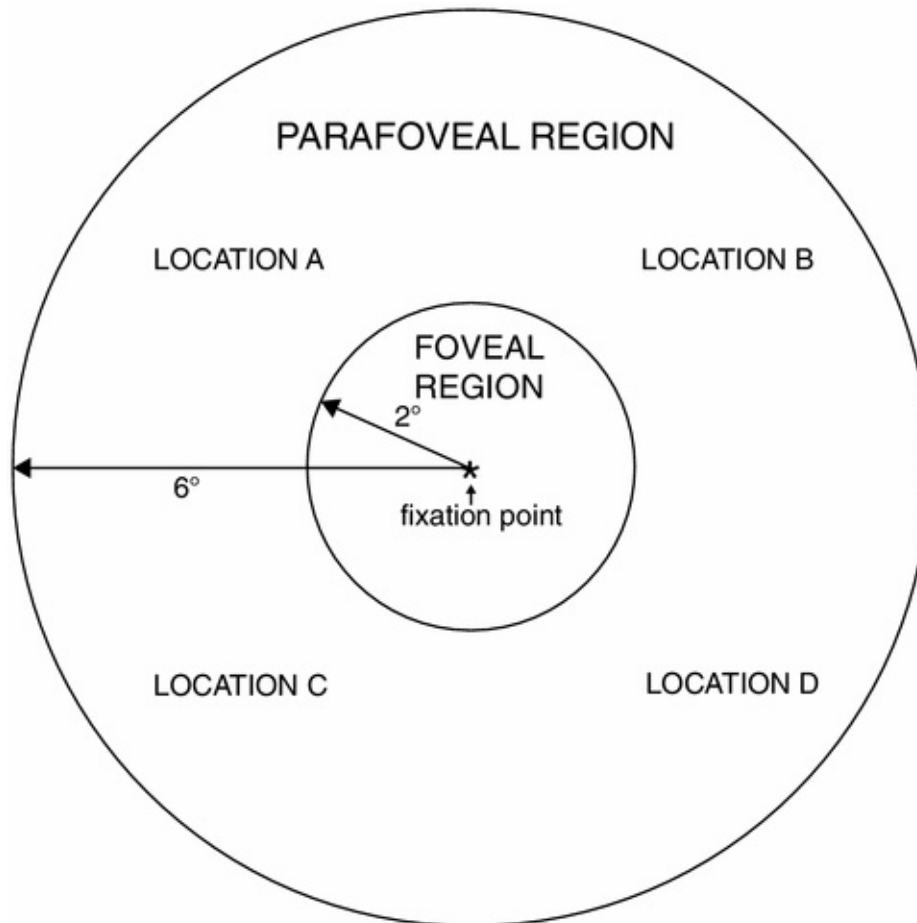
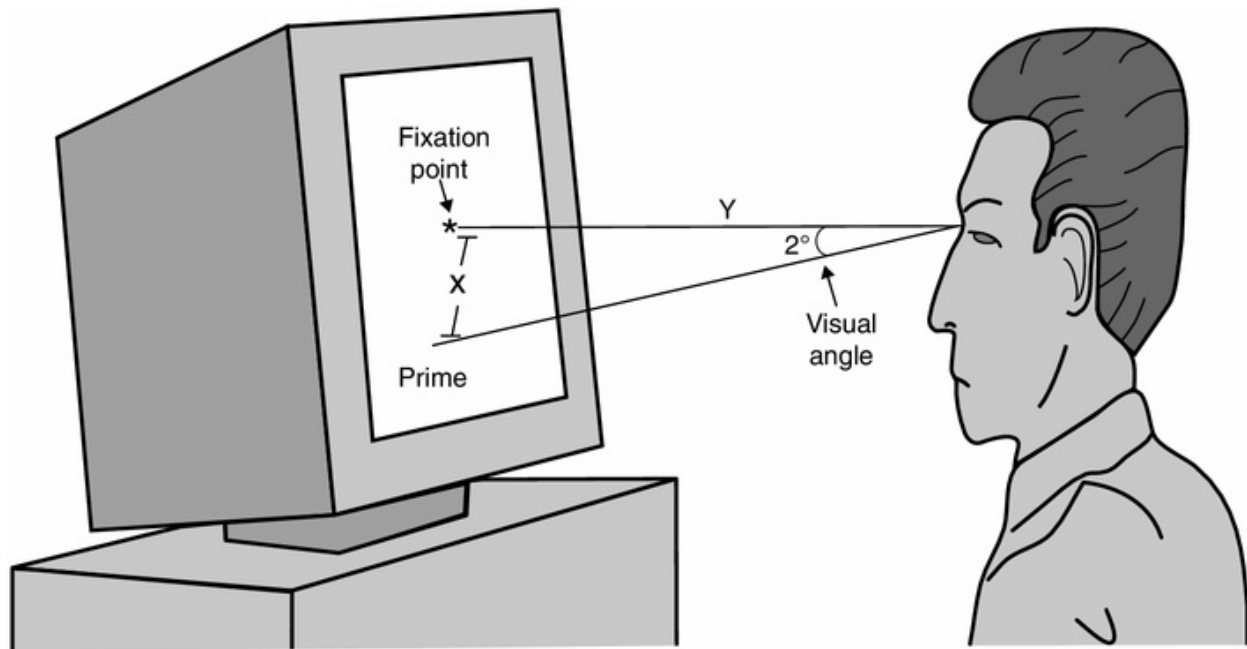


Figure 13.1. Subliminal priming: the foveal and parafoveal visual fields.

The parafoveal visual field extends beyond the foveal one, from about 2 to 6 degrees of visual angle. Determining the parafoveal area of computer display screens involves taking into account the distance between the participant's eyes and the screen; the farther away the participant is seated from the screen, the greater the area on the screen that falls in the foveal region (see [Figure 13.2](#) and Bargh et al., 1986 for details on calculating the visual angle).



$$Y = X/\tan(2^\circ)$$

Figure 13.2. Determining the parafoveal region of the computer display and calculating the required distance of the participant from the computer screen.

Information presented in the parafoveal region does not reach conscious awareness, at least as concerns its meaning or identity. One does become aware of movement and changes in this region, which automatically attract attention. However, parafoveal information is processed subconsciously to some extent. One can in this way “get away with” longer presentation durations with parafoveal compared with foveal presentation. The study of parafoveal processing has been a major topic in research on reading, specifically on one's ability to anticipate or “look ahead” in order to facilitate the fluent conscious processing of the material (Raynor, 1978).

Masking.

It is usually not sufficient to present a prime briefly and then remove it from the display. This is because the effective duration of the stimulus is longer than the actual duration, for two reasons. First, the decay rate of the medium in which the stimulus is electronically presented (this is not a problem for tachistoscopic display) is greater than zero. Older monitors (ca. 1980s) used a phosphor medium that sometimes took so long to decay that you could watch it happen.

More modern computer monitors have much faster decay rates, but it is important to look into this specification before purchasing equipment. The best of today's monitors have such fast decay rates that it is no longer a problem, at least for the kind of subliminal studies usually done in social psychology.

But even if one is using a tachistoscope or the best computer monitor on the market, one still needs to mask the stimulus. This brings us to the second reason why a stimulus duration could be longer than intended (and attain conscious awareness): because it tends to persist in the participant's visual iconic memory store for a time after it has physically disappeared from the display (see Sperling, 1960 for the first demonstration of visual iconic storage). To erase or overwrite the visual buffer so that the effective presentation duration of the prime is the same as its actual duration on the screen, a pattern mask should be presented at the same location, overwriting the prime on the display, and for as long as – and preferably longer than – the prime had been presented (Marcel, 1983; Turvey, 1973). A mask following the priming stimulus is called a backward mask. One can also use a forward mask that is presented immediately prior to the priming stimulus. This is particularly useful when the primes are pictures or photos rather than words, because the threshold for conscious awareness tends to be lower.

A pattern mask contains the same features as does the prime so that the same mental feature detectors are used in perceiving it. However, so as not to interfere with the effect of the prime, the pattern mask should not correspond to any higher-level meaning. Thus, for example, the primes in the Bargh *et al.* (1986) study were all words, and so the masking string (“XQFBZRMQWGBX”) was made up of the same features – that is, letters – but was not itself a word. In another example, Fitzsimons, Chartrand, and Fitzsimons (2008) used brand logos (such as the well-known Apple logo) as the subliminal priming stimuli in their studies, and so the mask was a square containing a pattern with the same colors as the logos (see Appendix C at the end of the chapter). In this way the same feature detectors are employed for prime and mask, disrupting the visual iconic storage.

With immediate pattern masking, the prime can be presented outside of awareness at durations of 15 ms or below for foveally presented faces (Bargh, Chen *et al.*, 1996, Experiment 3; Edwards, 1990) and schematic line-art renderings (i.e., cartoon-like drawings) of faces (Niedenthal, 1990).³ When parafoveal presentation is used instead of foveal, longer durations can be used: 60 ms in Bargh *et al.* (1986) and Chartrand and Bargh (1996), 125 ms in Erdley and D’Agostino (1988), and 90 ms in Bargh *et al.* (1995, Experiment 1).⁴

With parafoveal presentation it is important to ensure that the prime is really presented parafoveally – that is, that the participant's visual focus is on the desired fixation point. Only then can one be entirely sure that the prime was parafoveally and not foveally presented. For instance, if the parafoveal prime is always presented at the same point in time in a trial (say, 1 s after a warning signal), it can easily be anticipated, and the participant's attention can move away from the instructed fixation point to the location of the flash (making it phenomenally foveal regardless of the experimenter's intention).

To avoid that possibility, we have often inserted a random delay of 2–7 s between the trial warning signal and the presentation of the prime (Bargh & Pietromonaco, 1982; Bargh et al., 1986; Chartrand & Bargh, 1996). In addition, the prime was presented in one of four possible locations (“quadrants”) on the screen (all in the parafoveal region). Which one of these was used for a given trial was determined randomly by the computer, thus minimizing the possibility of anticipations by the participant.

Could the participant move his or her eyes quickly enough after the presentation of the prime to “catch” it before it is masked and thereby consciously see its content? The answer is no, as long as the parafoveal presentation is brief enough. The normal speed of saccadic jumps of the eye from one location to another is about 220 ms, by which time the presented prime and mask are long gone from the display. (There is some controversy over the existence of even faster saccadic jumps of 100 ms, called “express saccades” [Fischer & Weber, 1993], but if the parafoveal presentation is 60 ms or so, even these could not get to the presentation location in time.)

One way to ensure that the participant's attention is focused on the fixation point at the time of the parafoveal “flash” is to give him or her some task to perform involving stimuli presented at the fixation point. For instance, the participant could be asked to repeat out loud each of a series of digits presented at the fixation point, with the experimenter keeping track of correct performance. The prime could be presented immediately following the presentation of the final digit so that if the participant reported it correctly, visual attention could be safely assumed to not have been at the presentation location. (One could go further and vary the number of digits presented on each trial and in this way prevent the participant from anticipating the moment of prime presentation.) Another way to keeping participants’ attention on the fixation point is to require them to keep a running total of the numbers (single digit) that appear in the center of the screen (e.g., Fitzsimons et al., 2008). At the end of the procedure

they report the total sum, and if it is correct, it indicates that their attention was indeed on the fixation point.

Awareness Checks for Subliminal Priming Tasks.

With subliminal priming, one should probe for awareness of the relation between the priming and experimental tasks, just as with supraliminal priming. It is always possible for the participant to “get lucky” and happen to be looking right at the prime location at the moment it was presented, all of the experimenter's precautions notwithstanding. And it only takes conscious awareness of one prime to potentially alert the participant to the nature of the priming stimuli and consequently raise the specter of demand effects.

As an awareness check, the experimenter could follow up the experimental trials with a short representation of some of the original priming trials. The participant should be informed this time that words (or pictures) are being presented and to try to guess what they are. If the participant is not able to guess any of the words or identify the gist of the pictorial content, it is safe to say that subliminal presentation has been achieved. An even more conservative test would be to give participants the correct answer along with one or more distracter items prior to each trial of the awareness check and compare performance with that of a control group to which no actual primes are presented (Bargh et al., 1986).

Note in this regard that comparing performance to chance levels (e.g., 50% on two-item tests) is not an appropriate awareness check, because the particular distracters that are used can vary in how likely they are to be chosen given no primes. Factors such as word frequency or relevance to psychological issues (e.g., personality trait terms vs. vegetable names as distracters) play a role in the frequency with which both distracters and target primes are chosen (Fowler, Wolford, Slade, & Tassinari, 1981). Thus comparisons need to be made between the frequency with which the distracters are chosen in the prime and a no-prime control condition.

Our recommendation is to forego giving the participant options from which to choose and to base judgments of awareness on his or her ability to consciously report the prime stimulus after each trial of the awareness check task. Better than chance performance in selecting the correct item from a set of options could come from actual awareness, but it also could be the result of priming itself! (Indeed, such a result might well be expected on theoretical grounds.) If one instead uses any effect of the prime on task performance as the definition of

awareness, subliminal effects are defined out of existence (e.g., Holender, 1986) – and this would not seem to be a very interesting or productive route to take.

Supraliminal and Subliminal Priming Compared

This brings us to an important point about the role of awareness in priming effects. The same effects have been repeatedly obtained with subliminal and supraliminal priming manipulations alike: assimilation of ambiguous but relevant input into the primed category or activation of the primed goal. For example, in the subliminal priming studies described earlier (Bargh & Pietromonaco, 1982), the same assimilative priming effect was obtained as in the original Higgins *et al.* (1977) and Srull and Wyer (1979) studies that used supraliminal primes. Thus, awareness of the priming stimuli's presentation does not matter for the obtained effect.

However, awareness of the potential effect or influence of the priming events does matter. This may specifically become an issue when using supraliminal priming procedures. If the primes are very extreme exemplars of the category (e.g., Hitler and Dracula as primes for “hostile”; Herr, Sherman, & Fazio, 1984), they are especially memorable and likely to be used as a conscious standard of comparison subsequently. Target person Donald's refusal to pay his rent pales in comparison as an example of hostile behavior next to the exploits of Torquemada; if one has just read about the horrors of the Spanish Inquisition, one would probably see Donald as less, not more, hostile than otherwise. Strack and Hannover (1996) provided a thorough analysis of when such contrast effects are to be expected. The most important factor seems to be whether the priming event is still in conscious awareness (or working memory) at the later, critical moment (Lombardi, Higgins, & Bargh, 1987; Newman & Uleman, 1990).

So, if a person is aware of the relevance of the priming event to the later perception or judgment, there is an adjustment away from the presumed effect of that event (i.e., the person's “theory” of how they would have been influenced; Wegener & Petty, 1995; Wilson & Brekke, 1994). But in the usual case, in which one is not aware of the potential influence, bias in the direction of the primed representation occurs.

Clearly, then, what matters for the occurrence of unintended effects of the environment on one's thought, feeling, and behavior is not the lack of awareness of the occurrence of the event – which is how cognitive psychologists typically define unconscious influences (Greenwald 1992; Greenwald, Draine, & Abrams, 1996; Shevrin, 1992) – but rather a lack of awareness of the potential influence

of that event. One can be consciously aware of the event and still have it affect or even control one's thought or behavior. (For a vigorous historical debate as to one's degree of awareness of mental processes more generally, see Ericsson & Simon, 1980; Nisbett & Wilson, 1977.)

Strength of Priming Manipulations

In general, the more priming stimuli presented to the participant, the stronger the obtained priming effects. Srull and Wyer (1979) varied both the number of items in the scrambled sentence test (30 or 60) and the proportion of the items containing trait-relevant primes (20% or 80%). Both factors produced significant main effects, meaning that the more total primes and the greater the concentration of relevant primes within the task, the stronger the priming effects on impressions.

As a general rule, the scrambled sentence test or other supraliminal priming tasks – that is, tasks in which the individual is aware of the priming material – produce stronger priming effects than does subliminal priming. Activation from a conscious, intentional processing of the primes is stronger than subconscious activation, in the same way that increasing the brightness or duration of a stimulus on a tachistoscope eventually raises it from being invisible to visible. The stronger the activation of a concept, the greater its accessibility and likelihood of subsequent use (Higgins & King, 1981).

Moreover, the stronger the priming manipulation, the longer the priming effect lasts. Higgins, Bargh, and Lombardi (1985) explicitly tested a “synapse” model of concept accessibility in which the frequency of priming was pitted against recency of priming. Stimuli related to two different trait constructs (e.g., “adventurous” and “reckless”) were presented in a scrambled sentence test, but with one trait primed more frequently during the course of the task and the other primed more recently (i.e., on the final trial). Then, in the ostensibly unrelated task that followed, participants read about a target person who behaved in a way applicable to both primed concepts (e.g., sailing alone across the Atlantic). Participants’ impressions of the target person were more consistent with the evaluative implications of the recently primed trait if they were asked their opinion right after the priming task, but more consistent with the evaluative implications of the frequently primed trait if asked a few minutes later.

However, in the quest for a powerful priming effect, one must be careful not to overdo it. Great care should be taken when designing and conducting priming research in order to rule out active effects of the priming manipulations – the

most notorious of these being demand effects. A manipulation that is too heavy-handed is likely to tip off the participant as to the nature of the study, especially when they see the “Donald” story served up next in which the protagonist behaves somewhat in line with that trait.

Beyond Perception: Goal and Behavior Priming

A development in priming research that occurred in the 1990s is the opening up of the range of psychological phenomena that can be primed. For many years priming research focused exclusively on effects in perception and impression formation (see reviews in Bargh, 1994; Higgins, 1989; Wyer & Srull, 1989). Although some studies did employ a dependent variable that was not a judgment – for example, the participants’ behavior toward another person or toward an attitude object (Carver, Ganellen, Froming, & Chambers, 1983; Fazio, Chen, McDonel, & Sherman, 1982; Herr, 1986; Neuberg, 1988) – it was a priming effect on an evaluation or judgment that mediated the behavioral effect.

About 20 years ago researchers found that the same priming manipulations used to produce perceptual effects, such as the scrambled sentence test, produce behavioral or motivational effects as well, if that kind of dependent measure is employed instead. That is, it is possible to prime a behavioral tendency or prime a particular goal via the same manipulation (supraliminal or subliminal) originally employed to produce perceptual effects. For example, Bargh, Chen *et al.* (1996, Experiment 1) used a scrambled sentence test to activate the concept of rudeness or politeness and then waited to see if the participant would next interrupt an ongoing conversation. Those primed with rude stimuli were much more likely to interrupt (63%) than were nonprimed participants (38%), and those primed with politeness interrupted the least often of all (17%). Importantly, this effect was not mediated by the participants’ impressions of the experimenter, which suggested that it was attributable to a direct effect of priming on behavioral tendencies (as originally predicted from the theoretical position that there is a direct passive effect of perception on action; Prinz, 1990). This is in harmony with more recent neuroscience research on “mirror neurons” directly connecting perceptual and behavioral regions of the brain; *e.g.* Rizzolatti & Sinigaglia, 2008).

Motivations and goals can also be primed. In the original demonstrations by Bargh and Gollwitzer (1994; see also Bargh, Gollwitzer, Lee-Chai, Barndollar, & Troetschel, 2001), achievement or affiliation motives were activated by having participants first perform a “word search” task. Embedded in a matrix of

letters were words synonymous with one or the other motivation. Those primed with achievement worked harder and found more words in subsequent word search tasks compared with participants primed with affiliation, who were more concerned with interacting with the confederate than with working on the task.

The purpose of the Chartrand and Bargh (1996) studies was to show that primed information-processing goals operated the same way as did consciously and intentionally activated goals. Our first experiment used a scrambled sentence test to prime either the goal of forming an impression or of memorization (shown in [Appendix A](#) at the end of the chapter). Next, in an ostensibly unrelated second experiment, participants were presented with the set of social behaviors used in Hamilton, Katz, and Leirer (1980b). We obtained the same results as in the Hamilton *et al.* (1980b) study – higher free recall of the behaviors and a greater degree of thematic organization of them in memory in the impression than in the memory condition – even though we primed those goals instead of giving them to participants directly through experimental instructions. And in our second experiment, we replicated previous findings of on-line impression formation (Bargh & Thein, 1985; Hastie & Park, 1986) using subliminal priming of the impression goal instead of explicit conscious instructions to the participant to form an impression.

Since these original studies, there have been many further demonstrations of goal priming in adults and even infants, across a variety of goal domains (e.g., Dijksterhuis, Chartrand, & Aarts, 2007; Over & Carpenter, 2009), and impressive advances in understanding the mechanics through which unconscious goal pursuit operates (Aarts, Custers, & Marien, 2008).

What Have We Been Priming All These Years?

It is noteworthy that the same priming methods – such as the scrambled sentence test and subliminal prime presentation – produce motivational and behavioral as well as perceptual effects. The inescapable conclusion from this fact is that in a given experiment, a priming manipulation simultaneously produces all of these various effects. Just because the dependent variable of interest in a given study is, say, impressions of a target person, this does not mean that the only effect of the priming manipulation on the participant was on his or her social perception. If the experimenter had instead placed the same participant in a situation in which he or she could behave in line with the primed construct, behavioral effects would have been obtained.

Priming effects therefore occur and operate in parallel, just as do automatic

processing effects. Priming manipulations have more influences on the participants (and on people in real life) than happen to be measured by the experimenter. It is in our view one very important future direction for priming and automaticity research to sort out how these various simultaneous processes interact with one another (for an excellent example of such research, see Moskowitz, Gollwitzer, Wasel, & Schaal, 1999).

Mindset Priming

Mindset-priming studies, reviewed in this section, also prime motivations or processing goals but do so by having the participant first engage in that goal or intentionally use the mental procedure in question. Because priming involves active and intentional use of the procedure, and not just the passive activation of the goal concept, we consider mindset priming to be of a different variety than conceptual priming is. Mindset priming is characterized better as a carryover of an intentionally pursued goal or mental procedure to a new context. An act of conscious will on the part of the participant is required, unlike in conceptual priming.

As a result, there is a greater role played by intention and awareness in mindset priming, which makes studies using this technique more susceptible to demand effects. Nevertheless, it is sometimes more appropriate to use a carryover priming paradigm than a conceptual one. For instance, if the concept to be primed is too abstract or too procedural to prime in a scrambled sentence task or subliminal priming procedure, it might be more reasonable to use a carryover priming task. Moreover, it is a legitimate matter of interest whether intentional goal pursuit in one context influences the individual's decisions and behavior in subsequent contexts, without their awareness (or choosing) of this goal at the later point in time.

The original study of this kind was performed by Gollwitzer, Heckhausen, and Steller (1990). The participant was instructed to think about a personal problem in one of two ways: either to dwell on the pros and cons of a specific way to solve the problem (inducing a *deliberative* mindset) or to generate a specific detailed plan to accomplish an important personal life project (inducing an *implemental* mindset). (Control condition participants merely looked at a book of photographs during the same time period.) In the ostensibly unrelated second experiment, participants were given the first few lines of several novel “fairy tales” and were instructed to complete each tale. They could complete the story

any way they liked, but as predicted, those who had previously been given an implemental (action-oriented) mindset were more likely than the other participants to continue the story with what the protagonist actually did in order to accomplish a chosen goal, whereas those participants previously in a deliberative mindset more often wrote endings in which the protagonist considered and chose between various action alternatives. These findings suggested that the goal or mindset used in the first experiment continued to be active and operate in the second task, without participants being aware of or intentionally choosing this mode of thought while writing the story endings.

A second example of mindset priming comes from research by Bator and Cialdini (1995; see also Cialdini, 1994). In a first experiment, motivations to hold consistent beliefs (i.e., cognitive consistency) were primed in some participants. This was done in the following manner. Participants were told that they would be interacting with another person and then read an essay purportedly written by that person. The content of this essay either indicated that the other person very much valued consistency in beliefs and behavior, or it did not indicate this. Next, in what was presented as an unrelated experiment, all participants were asked to write an essay in favor of having comprehensive examinations instituted as a graduation requirement – something nearly all of these college students personally opposed. Participants wrote this counterattitudinal essay either under free-choice (i.e., they were asked to by the experimenter but could ostensibly say no) or no-choice (i.e., they were instructed to by the experimenter) conditions, following which they were asked for their own positions on the issue.

According to cognitive dissonance theory (e.g., Wicklund & Brehm, 1976), writing counterattitudinal essays under free-choice conditions should cause the participant to become more favorable toward the issue, compared with participants who felt they had no choice in writing the essay. However, Bator and Cialdini (1995) obtained this effect only for those participants whose consistency motivation had been primed. Participants in the control (not primed) condition held the same final position on the comprehensive exam issue regardless of whether they had written the essay under free-choice or no-choice conditions.

Unwanted Effects of Priming

Priming is an experimental sword that cuts both ways. That is, a participant's recent experience in an experimental setting will potentially affect his or her

subsequent responses whether or not such an effect was intended by the experimenter. Having participants complete questionnaires prior to another dependent measure can be a major source of unwanted priming effects (i.e., unless of course the experimenter has planned for and wants this influence). This is because in the course of the questionnaire the participant will consider and use concepts that then become more accessible and likely to be used, if relevant, in subsequent experimental tasks. This is especially a problem if the experimenter wishes to draw conclusions about the chronic or long-term nature of the effects found in the latter tasks, because the temporarily primed state of the influential concepts might have produced the effects instead.

This has now been demonstrated in several studies. When Skelton and Strohmets (1990) had some participants first rate a series of words on their health connotations, those participants subsequently reported having a greater number of health problems as measured by symptom checklists. Marks, Sinclair, and Wellens (1991) gave their depressed and nondepressed participants the Beck Depression Inventory (BDI) at the beginning of the experimental session and thereby produced different self-judgments compared with those of participants who had not earlier completed the BDI.

Any good experimental design is informed by a task analysis, in which the experimenter carefully considers how the various manipulations and tasks will affect the psychological state of the participant. Our advice is to include in such task analyses a consideration of how tasks positioned earlier in the experimental session could possibly, through conceptual or mindset priming, influence dependent measures positioned later in the session. A failure to do this yourself at the design stage runs the risk of having a reviewer or journal editor do it for you later on.

Demand Characteristics and Mindset Priming

Priming manipulations seek to activate concepts in one context to study the passive effects of this activation in a subsequent task. Conceptual priming produces this activation with a first task that is as different from the experimental task as possible, to show that it is the mere activation of the concept – not the source of or reason for the activation⁵ – that matters. Mindset priming, however, involves the active use of a certain way of thinking (at least vicariously) by the participant in the first experiment, which is then more likely than otherwise to be employed in the second task.

Because of this, one has to be much more worried about experimental demand as an explanation for mindset than for conceptual priming results. The skeptic could argue that by being told to deal with information in one way in the first task by the experimenter, the participant assumes that this is what he or she is supposed to do with the information presented in the second task. Extra care should be taken, therefore, to camouflage the relation between the two tasks as much as possible (e.g., by using different rooms and experimenters for them) and to probe carefully for awareness of the relation between the two tasks (see [Appendix B](#) at the end of the chapter).

Automaticity Research Techniques

As discussed in the section on the history of automaticity, there never was such a thing as a single type of processing, called “automatic,” that could be studied with just a single paradigm or methodology. Instead, different paradigms and tests have evolved to study the separate qualities of the not-conscious processes that are grouped under the umbrella category of “automaticity.” These separate qualities are (a) whether the individual is aware of the operation of the process, (b) whether the process is efficient, (c) whether it is unintentional, and (d) whether the individual can control the process. Although tests of awareness of a process have already been discussed in the Subliminal Priming section (see also Murphy, Monahan, & Zajonc, [1995](#); Murphy & Zajonc, [1993](#)), there are distinct methods of testing for the presence of each of the other three qualities of automaticity.

Efficiency

Efficiency in processing is important to study because there are usually many demands on our limited attention, or working memory, at any given moment. Processes that do not require much, if any, conscious attention to operate will therefore have an advantage under these busy circumstances. They will occur more consistently over time in a given situation and constitute the default set of reactions to most occasions (Bargh, [1997](#); Brewer, [1988](#); Fiske & Neuberg, [1990](#); Gilbert & Osborne, [1989](#); Gilbert, Pelham, & Krull, [1988](#); Rothbart, [1981](#)). Therefore, it is important for us as social psychologists, as we are in the business of studying the individual's general and typical reactions to situations, to study the efficiency of any process on which we are focused. As Langer ([1978](#)) noted more than three decades ago, we as researchers are not on sure

footing when we generalize to the noisy real world the results of laboratory studies in which our participants are given plenty of time and nothing else to do while the critical phenomenon is being scrutinized.

Many models of stereotyping (e.g., Devine, 1989), causal attribution (e.g., Gilbert, 1989; Trope, 1986), and impression formation (e.g., Bargh & Thein, 1985; Brewer, 1988; Fiske & Neuberg, 1990) posit two stages of processing. One is the default and is described as very efficient; the second stage is more effortful and can only occur if the person has the time, attention, and motivation to use them. We leave a consideration of the motivational variable to the next section on unintentional processing. Efficiency per se allows a process to operate in both of two “real-world” conditions of information overload: when there is no time to consider and integrate the various available sources of information (such as a rapid stream of behavior, emotional reactions, and so forth during impression formation) and when one's current goals and purposes take attention away from what is going on in the environment.

The attentional demands made by a mental process can be measured directly, typically through reaction time techniques, or the attentional demands of a task can be manipulated to assess if performance is affected. Either method can yield information about the efficiency of the underlying process – its ability to operate under conditions of scarce attentional resources.

Measurement of Efficiency.

It is possible to measure the efficiency of a mental process in terms of how much time a person requires to engage in it. Smith (1994) and his colleagues performed a series of studies demonstrating the development of procedural automaticity in the domain of social judgments. In their paradigm, participants judge whether each of a series of behaviors is or is not an instance of a particular personality trait. The speed with which this yes-no decision is made is measured (in milliseconds). It is shown that the time to make these trait categorizations of behaviors decreases with practice, demonstrating an increase in procedural efficiency or automaticity. This *proceduralization* has two components: a general component in that judging behaviors with regard to a particular trait (e.g., kindness) becomes faster even with novel behaviors (not judged previously), and a specific component in that the same behavior judged in terms of the same trait is done still more efficiently (Smith, Branscombe, & Bormann, 1988; Smith & Lerner, 1986; Smith, Stewart, & Buttram, 1992). The speed-up with practice was also found to follow the same inverse power function that

characterized nonsocial mental process proceduralization (e.g., Anderson, 1982).

Although it is true that the more efficient a mental process, the less time it requires to run to completion (because conscious attention can only be deployed over time; Logan, 1980), the converse does not follow. That is, one cannot directly infer from the amount of time participants take to make a judgment or decision, for example, how efficient or automatic it is. This is because other factors influence and contribute to response times besides the procedural efficiency of an underlying process – most notably, strategic self-presentation. We treat this issue in more detail later (see “Some issues regarding the use of reaction times as a dependent variable”). The research of Smith and colleagues is a good example of a paradigm in which one is able to draw conclusions about underlying procedural efficiency from raw response times, because the same behaviors are being judged by the same participants repeatedly, in a within-subjects design. Thus, other influences on response times, such as how long it takes the participant to read the behavior, are held constant across trials. Moreover, because the participant is not making self-referential decisions about the behaviors, no self-presentational strategy is likely to be operative.

An interesting variant of measuring efficiency through response times can be found in the work by Macrae and his colleagues (e.g., Macrae, Milne, & Bodenhausen, 1994) on the automaticity of stereotype activation. Instead of measuring latencies in the primary task given to participants, as in the research by Smith and colleagues, these researchers made use of a dual-task procedure to measure response times to a secondary task. Participants were instructed to monitor a tape-recorded informational passage about Indonesia at the same time as viewing information on a computer screen about a target individual and forming an impression of him. Some participants were given a stereotypic label (e.g., “skinhead”) about the target. The interesting twist on the usual dual-task paradigm was that it was performance on the secondary task that was the dependent variable of interest. It was found that performance on the prose-monitoring task, as tapped by later memory for it, was better if stereotype-relevant information had been presented in the course of the impression-formation task. This confirmed the authors’ hypothesis concerning the efficiency with which stereotypes process relevant information.

Manipulation of Attentional Demands.

One can also assess efficiency of a process by manipulating the attentional demands of a task, to see if this changes task performance. To the extent that it

does, attention is needed for the task; to the extent that it does not, the process is unaffected by attentional shortage and is thus quite efficient. Accordingly, laboratory manipulations of these conditions either present information very rapidly (information overload) or give the participant a secondary task to “load” attentional capacity (what Gilbert et al., 1988 termed “cognitive busyness”).

As an example, Bargh and Thein (1985) conducted a person memory study in which a series of 24 behaviors related to the trait of honesty (either honest, dishonest, or neutral behaviors as developed by Hastie & Kumar, 1979) was presented one at a time on a computer screen, and participants were instructed to form an impression of the target person who had performed these behaviors. In one condition, participants could read each behavior at their leisure, pressing the space bar to move on to the next behavior. (This technique had the additional advantage of allowing us to measure how much attention and consideration were given to the various types of information, as operationalized by looking time; see also Fiske, 1980.) But in the rapid-paced condition, each behavior was presented for only 1 s, with a 1-s pause before the next behavior came on the screen. This was just enough time for the participant to read each behavior one time through, preventing any further conscious deliberation about a given piece of behavioral information or its integration with others to form a coherent impression on-line. Results confirmed that this manipulation prevented participants from forming an impression on-line (at the time of reading the behaviors), forcing them to do so only later, based on those behaviors they could recall.

Gilbert and Osborne (1989) used a variation of the memory-load technique that has become a popular methodology because of its simplicity and effectiveness (e.g., Macrae, Hewstone, & Griffiths, 1993; Wegner & Erber, 1992). They gave participants a single eight-digit number to remember throughout the entire time that the critical person information was presented (via a videotape) and only after all of the information had been presented did they ask participants to repeat the number back. They found predicted differences in attributions and judgments as a function of this cognitive load manipulation; the memory load prevented participants from being able to take situational influences into account in their behavioral attributions. Thus participants were more likely to make dispositional attributions under memory load even when clear situational forces were operating to constrain or shape behavior.

It is very important when conducting dual-task studies of this sort to make sure of some things. First, the “load” task must be sufficiently attention-demanding so that little attention remains with which to perform the primary

task. Imagine, for example, if in the aforementioned studies participants had been given a one-or two-digit number to remember instead of a six-or eight-digit number. Judgment latencies, or the type of attribution made across all experimental conditions, would likely not be any different from the nonload conditions, but it would be erroneous to conclude from this that making judgments never requires any attention or that situational attributions are made automatically. We would have the usual interpretational problem of null results. Thus it is best to include in the design conditions under which one does expect the memory load to have an effect, so that one is confident that the load was sufficiently strong to affect the dependent variables in conditions where it is theoretically expected to do so, whereas not affecting them in the conditions where one's theory predicts relatively attention-free task performance.

One difficulty with having participants remember the same digit string throughout the experiment is that they learn it – that is, they store it in long-term memory, so that they may not need to keep rehearsing it in short-term memory. If a participant successfully learns the string – and in the Gilbert and other studies using this procedure the participant is given a minute or so before the experimental task starts to rehearse the number – then clearly the demands on his or her attention capacity would not have increased to any significant degree.

To show that the load manipulation is strong enough, it should be shown independently in a manipulation check to be of sufficient difficulty that participants do not perform it perfectly. In other words, it is good to show that they make errors in reporting the material they were to hold in memory (if that is their secondary task). But if they make too many errors, one cannot be sure if they were trying hard enough to perform that secondary task. One strategy the participant might take for coping with the attention load situation might be to disregard one of the two tasks and focus on one exclusively, to the detriment of performance on the other. If the participant adopts this strategy, no valid conclusions can be drawn about the attention demands of the primary task.

And so we as experimenters want the participant to make some errors, but not too many. The solution to this problem adopted by Gilbert *et al.* (1988) and Gilbert and Hixon (1991) was to omit data from participants if they did not report at least half of the digit string correctly. Another possibility is to include, as either a between-subjects (separate set of participants) or within-subjects (additional repeated measure on the same participants) control condition, an even stronger load manipulation. If this additional condition produces the same results as the original load condition, then the latter was most likely completely

loading the participants' working memory; if the results differ, then the original load manipulation was not completely using available attentional resources.

It is also possible to test out one's load manipulation via a pretest in which participants are given a task or manipulation known to require conscious effort; the no-load condition should replicate previous findings on this task, but the load manipulation should knock out this effect. This load manipulation check technique was employed by Moskowitz *et al.* (1999) using outcome dependency as the test manipulation; the load effect successfully eliminated the usual effect of outcome dependency (i.e., to increase effortful scrutiny of the target person).

In dual-task paradigms, it is important that the participant consider the experimental task to be the primary one – that is, the more important of the two (Kantowitz, 1974). To assess the attentional demands of a primary task, everything should be kept as similar as possible about that task in the load and nonload conditions, other than the load itself. If in the load condition the participant believes the primary task to not be as important, and so is not as motivated to perform it compared with participants in the load condition, more than just the attention demands have changed to potentially affect the dependent measures. Thus participants should not be told that the tasks are equally important but instead that – although it is important for them to perform both tasks, not just one or the other – the focal (judgment, attribution, etc.) task is the crucial one for the experiment.

An interesting variant of the memory load procedure was introduced by Tice, Butler, Muraven, and Stillwell (1995). They were interested in the relative automaticity of self-presentational strategies to friends versus to strangers. The content of self-presentations to friends was found to be more modest than those to strangers. But the automaticity of these self-presentational strategies was assessed by the participants' subsequent recall of the interaction. The authors reasoned that the more that attention is focused inward, on one's own interaction performance, the less should be available for external events. Consequently, one's later memory for those events will be poorer. (This phenomenon was originally known as the “next-in-line effect”; Brenner, 1973). Tice *et al.* (1995) used this fact to measure the ease or relative automaticity of the different self-presentational strategies. As predicted, when participants were instructed to engage in their natural tendencies – to be modest with friends and self-enhancing with strangers – their later recall of the interaction was better than if they had been instructed to engage in the contrary strategy (i.e., self-enhancement with friends and modesty with strangers).

Unintended Processing Effects

A major source of unintended effects on thinking, feeling, and doing is automatic associative connections in memory. If the (intended or unintended) activation of representation “A” then proceeds to activate representation “B” automatically, without any conscious intent or awareness involved, this latter representation can have an unintended effect on judgments, evaluations, and behavior. For example, Devine (1989) designed her study to show that (white) participants “went beyond the information given” in their stereotypic assumptions by priming them with some aspects of the African-American stereotype, but not directly with “hostility,” which is also an element of that stereotype. The priming manipulation nevertheless did influence subsequent judgments about a target person's hostility – an effect that could only have occurred if hostility had been activated unintentionally because of the automatic spread of activation within the stereotype. Bargh *et al.* (1995) showed how the activation of the concept of power spread automatically to the concept of sexual relations for those likely to sexually harass or aggress, as indicated by their greater attraction to a female confederate after only power, not sex, was primed.

There are two major ways of establishing the existence of such automatic connections: through analyses of output order in free-recall memory measures (“clustering”) and through sequential priming techniques.

Clustering Measures of Memory Organization.

Free-recall measures of memory can be utilized to get at the underlying structure and organization of memory. The guiding logic here is that the order in which participants remember and hence write down what they remember about a person or event reflects the way it has been encoded in memory. The connections formed between the elements of the person or event memory help determine what is most easily recalled later on. Given that judgments and decisions are often made based on what is later most easily recalled from memory about the person or event (Hastie & Park, 1986), the organization of material in memory can later determine, in a passive way, the outcome of those judgments.

Before one can examine clustering, free-recall protocols must first be coded for whether each item written by the participant should be considered “correct.” Whereas the appropriate unit of analysis (i.e., what is coded as correct or incorrect) is clear with single-word recall paradigms, it is not so clear when the

stimulus materials involve behavioral phrases or prose paragraphs. Although either a strict “verbatim” criterion or a more lenient “general meaning” or “gist” criterion may be used in these cases, researchers have normally not found significant differences in their results based on the use of these different criteria. Many end up basing their final analyses solely on the leniently scored “gist” protocols, in which an item is scored as correct if it captures the primary concept or meaning expressed in the original item (Chartrand & Bargh, 1996; Hamilton et al., 1980a, 1980b). However, researchers should choose the criterion most appropriate for their particular study based on whether verbatim recall is theoretically necessary to show or not.

A related issue concerns “intrusions” in free recall, which are items “recalled” by participants that were not present in the original stimulus material (see Srull, 1984 for an in-depth discussion of intrusions). Because intrusion rates may vary across experimental conditions in a systematic way, they should be analyzed and reported by researchers. It is possible to use intrusions in free recall as an indication of information “added in” to a memory by the schema or stereotype used to encode the original information, but as intrusions in free recall are typically rare, such studies have mainly used recognition memory tests in which “hit rates” (yes responses to actually presented items) and “false alarm rates” (yes responses to test foils that had never been presented) can be compared to separate out accurate retrieval from guessing biases (see Grier, 1971; Srull, 1984; Wyer & Gordon, 1982).

The most common method of determining the amount of clustering in free-recall protocols is to use one of various objective clustering techniques, in which the conceptual categories organizing the information are specified a priori by the experimenter. Many different clustering methods exist, each with its own equation that yields an overall clustering “score” for each recall protocol. One of the most widely used measures of category clustering in free recall, the Bousfield and Bousfield's (1966) deviation (BBD) measure, was one of the first to be developed. Essentially, this measure is a ratio of observed category repetitions to the number of such repetitions expected on the basis of chance.

One limitation of the BBD is that there is no fixed upper bound; a positive score indicates clustering above chance, but it is impossible to determine whether the score reflects perfect or less than perfect clustering. Specifically, the score for perfect or maximum possible clustering changes with the number of categories that the participant recalls and with the distribution of the total items recalled across categories. Furthermore, the BBD is affected by the total number

of items recalled. Finally, because it does not reflect a proportion of actual to total category repetitions above chance, it is difficult to make comparisons between experiments or between participants.

Alternative clustering measures do exist, however, such as the modified ratio of repetition (MRR; Bower, Lesgold, & Tieman, 1969), the clustering (C) index, and the deviation (D) index (Dalrymple-Alford, 1970). Robertson's (1995) model-based measure of clustering, $\kappa\alpha$, is highly related to the clustering index, but requires an iterative procedure to calculate its value. It has the advantage of placing more weight on those repetitions occurring at the beginning of the recall list, less to those in the middle of the recall list, and no weight to any repetitions occurring at the end of the recall list. (Also see Robertson, 1995 for a model of recall order that incorporates the clustering information with the serial order in which they are recalled and any interaction between presentation order effects and clustering.)

Many researchers have argued that the adjusted-ratio-of-clustering (ARC) index developed by Roenker, Thompson, and Brown (1971) is the best overall measure (Murphy, 1979; Ostrom, Pryor, & Simpson, 1981; Srull, 1984; Wyer & Gordon, 1982). Unlike many of the alternative measures, ARC yields a clustering score ranging from 0, indicating no clustering beyond what would be expected by chance, to 1, indicating perfect clustering. Moreover, it corrects for different numbers of categories that are presented as well as the number of categories recalled. Finally, ARC appears to be the least confounded with extraneous factors (Murphy, 1979). The computational formula for ARC is

$$ARC = \frac{R - E(R)}{N - K - (R)}, \quad (13.1)$$

where R = number of observed category repetitions, N = total number of all items recalled, K = number of conceptual categories represented in the presentation list, and $E(R)$ = expected number of category repetitions, $(\sum m(I)^2/N) - 1$, where m is the number of items from category I that are recalled.

Although researchers should choose clustering measures carefully, it should be noted that the various formulas are often highly intercorrelated. For instance, Hamilton *et al.* (1980b) and Chartrand and Bargh (1996) used both the BBD and ARC measures in their analyses of clustering and found the same pattern of means with both indices.

In addition to these popular objective clustering measures there are some alternative techniques for recall output analysis. One of these involves calculating conditional probabilities and is best exemplified by Srull's research on person memory (Srull, 1981; Srull, Lichtenstein, & Rothbart, 1985; Srull & Wyer, 1989). Participants were presented with a series of behaviors by a target person and instructed to form an impression of him or her. Most of the behaviors were consistent with a certain personality trait (e.g., honest), but a minority were inconsistent (e.g., dishonest) or unrelated to the trait in question. By examining the order in which the behaviors could later be recalled and calculating conditional probabilities of recalling one type (e.g., inconsistent), given that the same or another type (e.g., consistent) had just previously been recalled, Srull and his colleagues constructed a sophisticated process model of impression formation.

This model could make accurate, detailed predictions about how people give consideration to unexpected, impression-inconsistent information and attempt to reconcile and integrate these behaviors into an overall, coherent impression of the target. These predictions were generated from a model of associative structure, deduced backward from a fine-grained analysis of recall output order, tracing the mental route participants took to retrieve each target behavior. Importantly, calculating conditional probabilities was the more appropriate method of analysis in these studies, as levels of category clustering by the objective measures were at near-chance levels. Yet there did exist a highly systematic nature to the order of items recalled that was uncovered using this different technique.

Sequential Priming Techniques.

The sequential priming task permits conclusions about the automaticity of associative connections between memory representations. By varying the time delay between the presentation of a prime stimulus and of a target stimulus, and assessing the effect of the prime on responses to the target under these different time gaps, inferences can be drawn as to whether the effect was immediate and automatic or conscious and strategic. Essentially, if presentation of the prime affects responses to the target at time gaps too short for temporary, strategic responses to have been responsible, then the prime and target concepts can be said to be structurally associated in long-term memory. Accordingly, sequential priming tasks have become one of the most widely used experimental techniques in social psychological research on memory structure and automaticity.

Associative network theory (e.g., Anderson & Bower, 1973; Srull, 1981; Wyer & Carlston, 1979) holds that memory consists of interconnected nodes, with activation spreading automatically from one node to another. Activation will only spread if there is an associative link that has been formed, and the stronger the association, the more and faster the activation will spread to the related node. Early experiments testing associative network theory showed that responses to a target item (e.g., NURSE) were faster if an associated node (e.g., DOCTOR) had just been activated (Meyer & Schvaneveldt, 1971). Presumably, activation had spread from the node representing the prime to that representing the target so that when the target was presented, that location was already activated and so required less time to be activated in the response process.

Posner and Snyder (1975) added a strategic mode or component to spreading activation theory. They held that automatic activation effects were the default, but could be overruled by a current goal or strategy in the task if sufficient time were allowed for this attention-demanding strategy to operate. Automatic sequential priming effects for prime-target pairs such as doctor-nurse or sun-moon were relatively fast, occurring in 300 ms or less. Temporary strategic effects, on the other hand, take longer to develop because they require attentional (effortful) resources that take time to accrue (Logan, 1980). However, if there is attentional capacity and sufficient time, strategic expectancies are capable of inhibiting and overruling the automatic activation (see also Shallice, 1972).

Neely (1977) tested this model by varying the amount of time between the prime presentation onset and the target presentation onset, known as the stimulus onset asynchrony or SOA. In each trial, a prime appeared in the center of the display for a certain amount of time, then was erased, and the target word was presented at the same location. Target words were members of the category BODY (i.e., parts of the body such as heart or leg) or the category FURNITURE (e.g., chair, table), or were non-words (e.g., trone). The prime stimulus was either the word BODY or the word FURNITURE. The participants' task was lexical decision, in which they were to respond whether a target was a word or a non-word as quickly as they could.

A key element of Neely's (1977) design was to vary the delay between prime and target presentation. With brief delays (e.g., 250 ms), only automatic effects should be able to occur; thus, the prime BODY should facilitate (speed up) responses to names of parts of the body (and likewise for FURNITURE and names of pieces of furniture) because strong, automatic connections are assumed to exist between these target concepts and their higher-order category concept.

Only with longer delays (e.g., 750 ms) should strategic conscious expectations be able to influence responses. In the critical experimental condition, participants had a conscious expectancy for the opposite of the semantically consistent prime-target combination. In other words, they expected the BODY prime to be followed by names of pieces of furniture and for FURNITURE to be followed by names of body parts. However, the automatic effect would remain the same as always, as it reflects long-term associations and cannot flexibly adapt to temporarily altered circumstances. In line with the Posner-Snyder model, Neely (1977) found that under these conditions the category-name primes continued to facilitate responses to members of that category under the short prime-target delay conditions, but that under the longer prime-target delay, category-name primes facilitated responses to members of the alternative category.

The sequential priming paradigm used by Meyer and Schvaneveldt (1971) and Neely (1977) has been employed increasingly to study social psychological phenomena. Fazio *et al.* (1986) based their original study of automatic attitude activation on the Neely's (1977) paradigm. The names of various attitude objects (e.g., basketball, Reagan, ice cream) were presented as primes, and positive and negative adjectives (e.g., beautiful, terrible) appeared as targets. The SOA between prime and target was also varied, either 300 or 1,000 ms. Instead of the lexical decision task used by previous researchers, Fazio *et al.* (1986) instructed their participants to evaluate the target adjective as quickly as they could on each trial, by pressing one of two buttons, labeled “good” and “bad” (see Figure 13.3).

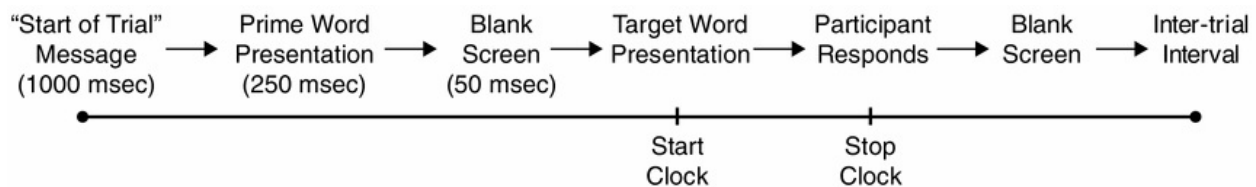


Figure 13.3. The sequential priming paradigm.

Results showed a reliable effect of at least some sets of attitude object primes on latencies to evaluate the target adjectives, with participants faster to respond when prime and target were of the same, rather than the opposite, valence. Importantly, this effect was found only for the short and not the long SOA conditions. For the attitude objects to affect the target evaluations at such short SOAs, the attitude object prime had to have activated its own evaluation before the target was presented – that is, within 300 ms – too quickly to have been the product of some conscious and intentional process. Moreover, the effect did not

occur when participants did have enough time (i.e., the 1,000 SOA condition) to prepare, strategically, a response based on the prime valence. This was presumably because prime valence was not diagnostic as to the valence of the target that followed; positive primes were followed half of the time by positive and half by negative primes, and the same was true for negative primes.

The outcome of these and subsequent studies on automatic attitude activation (Bargh, Chaiken, Govender, & Pratto, 1992; Bargh, Chaiken, Raymond, & Hymes, 1996; Duckworth, Bargh, Garcia, & Chaiken, 2002; Roskos-Ewoldsen & Fazio, 1992) have been uniformly consistent with the hypothesis that attitude objects immediately and automatically activate their associated evaluations in memory.⁶ Because the evaluations are made so quickly and without conscious intention, many researchers have now made use of the paradigm to investigate social attitudes that people are reticent to admit, such as stereotypic or negative views of social groups (e.g., Perdue et al., 1990). Because the dependent measure is the latency to respond while performing an innocuous task, there is no way for participants to strategically respond in a way that hides these automatic evaluations.

Fazio, Jackson, Dunton, and Williams (1995) used the automatic attitude effect itself as a predictor of prejudicial behavior. By assessing the degree to which African-American faces primed responses to negative adjectives and slowed down responses to positive adjectives, a measure of the participants' implicit stereotypic beliefs could be constructed unobtrusively. This measure was found to predict the negativity of the participants' behavioral reactions to an African-American experimenter, whereas a self-report measure of racial attitudes did not.

Although the evaluation task has become a popular one to use in sequential priming paradigms, a pronunciation task may often be preferable. Because the purpose in using the sequential priming paradigm is to establish the unintentional and immediate activation of social concepts and evaluations, conscious and intentional strategies on the part of participants should be eliminated from the paradigm as much as possible. Having participants intentionally evaluate adjectives in the test of attitude automaticity, for example, was problematic for drawing conclusions about the goal-independence or unintentionality of the effect. Participants were consciously thinking in terms of evaluation and were trying to evaluate the target adjectives – would the effect occur when this goal of evaluation was not currently operating? By having participants pronounce as quickly as possible rather than evaluate the targets, it

was shown that the effect did not depend on the conscious goal of evaluation (Bargh, Chaiken et al., 1996).

Balota and Lorch (1986) showed that the pronunciation task has advantages even over some apparently strategy-free tasks, such as lexical decision. For one thing, a researcher usually has to discard half of the data gathered in a lexical decision task because the responses to the non-word trials are not of theoretical meaning or importance. For another, lexical decision still involves a decision (i.e., word or non-word) about the stimulus, and this increases the time needed to respond and also the variance resulting from individual differences in the judgment process. In line with these reasons, Balota and Lorch found that pronunciation was a more sensitive measure of spreading activation than was lexical decision.

The sequential priming paradigm has the potential for illuminating many of the important situational effects that are at the heart of traditional social psychology (e.g., Ross & Nisbett, 1991). Instead of restricting ourselves to tracing the strong associative connections between internal abstract concepts, such as between elements in a stereotype, or between an object and its attitude, one can examine the immediate and unintentional reactions to social situations. This is quite simply accomplished by having the priming stimuli related to the situational features. A first attempt at extending the paradigm to situation-concept relations was successful in demonstrating automatic sexually related cognitions as a result of priming the situational feature of having power (Bargh, Raymond, Pryor, & Strack, 1995). Participants identified as likely to be sexual harassers or aggressors showed the sequential priming effect of power on sexually related stimuli in a pronunciation task, and in a second experiment were more attracted to a female confederate (compared with other participants) if the concept of power had been primed. Thus the sequential priming paradigm would seem to have great promise for investigating other automatic effects of situations, as well as individual differences in these reactions.

Some Issues Concerning the Use of Response Latencies as a Dependent Variable

The key dependent variable in the automatic evaluation studies, as in many other lines of social cognition research (e.g., the content of the self-schema; see Markus, 1977), is the speed with which a response can be made to the target stimulus. Response latencies can be very informative as to the accessibility and automaticity of concept activation and as to the automaticity of connections

between two concepts (i.e., prime and target stimuli), but there are two important caveats to keep in mind.

First of all, there are usually more components to a response latency than just the one that is of experimental interest. This is true for evaluation, lexical decision, and even pronunciation tasks. Take, for example, the operational definition of attitude strength in terms of latency of responding “good” or “bad” to the name of the attitude object (Fazio et al., 1986). The shorter this latency, the stronger the corresponding attitude was considered to be. However, many other factors influence the latency to respond to a given attitude stimulus, such as word length (it takes more time to read longer words) and word frequency, to name a few. These theoretically uninteresting features of the stimuli proved to be significantly correlated with evaluation latencies in further studies (Bargh et al., 1992). If one uses simple latencies alone, as if the only influence on them were attitude strength, one ends up making some erroneous inferences (e.g., concluding attitudes toward gum are stronger in general than attitudes toward abortion).

Perhaps more important, conscious response strategies can influence response latencies, especially those resulting from evaluation tasks. It should be noted, however, that researchers can avoid this particular problem by employing a pronunciation task for the sequential priming procedure, because pronunciation tasks are not as susceptible to response biases as are other tasks.

Rogers (1974) was the first to analyze response latencies to trait terms in self-judgment tasks in terms of both the degree to which the concept was part of the self-concept and in terms of the participant's strategy in answering the questions. This distinction between the actual latency component of interest and mere response strategy was an important one to make. One common strategy is positive self-presentation, causing fast latencies when saying no to negative items and yes to positive items, rather than vice versa. This result could occur either because self-concepts are generally positive or because the participant has adopted a strategy of basing his or her response not on the true self-concept but on merely matching the response to the positivity or negativity of the item.

Because such response strategies are effortful and require attentional resources, we recommend separating the activation and strategic components by loading attention with a secondary task to see if latencies are affected by the load. To the extent that the latencies are not affected by the load manipulation, this signals the true automaticity or chronic accessibility of the judgment process or underlying mental representation; to the extent memory load increases the

latency of response to that item, it can be concluded that the concept could only be responded to effortfully. Without assessing latencies under attentional load, the role played by response strategies remains unclear.

The second important caveat, which holds for all types of sequential priming tasks, is that the distribution of response latencies is typically positively skewed, in that they are constrained at the fast end and not at the slow end. This means a transformation must be carried out to normalize the distribution. There are a variety of possible transformations, such as taking the natural logarithm or the reciprocal of the raw latency. The natural logarithm is a milder transformation, whereas taking the reciprocal is somewhat stronger in that it alters the original distribution to a greater extent. The question of which of transformation should be used has been a matter of some debate. Fazio (1990) recommended the reciprocal transformation, but Winer (1971) argued against this as too strongly altering the underlying distribution, recommending instead the natural logarithm. (See also Box, Hunter, & Hunter, 1978 for a comparison of the effects of different transformations.) Perhaps the most reasonable method is to try several transformations (moving from mildest to strongest), examine their relative success in removing the positive skew, and then choose accordingly. Different sorts of tasks may have varying degrees of positive skew associated with them, and one wants to pick the transformation that does the best job in each specific case.

Along with distribution transformation comes the issue of what to do with outliers. These are very long latencies that can greatly affect the means and thus the outcome and conclusions from the experiment.⁷ It is usual and accepted practice to trim outliers to remove this distorting influence on the results. Some rules of thumb can be suggested.⁸

First of all, the same policy of trimming (and for that matter, of transformation) should be used in all of one's experiments as a matter of course.

Second, common sense as to what is a reasonable response latency for the task at hand should play a role in determining whether a long response is a true response or an error. For instance, if the task is merely to pronounce each stimulus word as quickly as possible after presentation, latencies of 1.5 or 2 s or longer would seem to indicate either an equipment error (e.g., the participant spoke too softly for the microphone to pick up the response) or a failure to follow instructions. But the same latency if the task is to say whether an adjective describes oneself is quite reasonable; it may easily take this long for the person to decide.

Latencies that are too fast to have been reasonable responses should also be trimmed; these are almost always anticipations and not true responses. Typically, latencies shorter than 300 ms are trimmed (and these are usually quite rare) for this reason. (Even the National Basketball Association endorses this 300-ms “minimal response time” rule – if less than 0.3 of a second remains on the game clock, no shot is allowed to count after play is resumed, as it is deemed impossible to get one off in this short a time.)

Third, only truly extreme latencies should be trimmed – for example, those that are more than three standard deviations above the mean (as in Blair & Banaji, 1996), or only the most extreme 2% of all responses. In the typical automatic attitude experiment, for instance, only between 1% and 2% of responses are trimmed.

Finally, it should be established that the deleted reaction times are equally distributed across conditions. If a disproportionate number of them fall in a given condition or subset of conditions, this implies that they are not random events or errors, but rather systematic effects of the experimental manipulations.

Because of the usual and recognized need in response latency research to trim and transform the data, it is more important than usual to earn and keep the trust of the consumers of your research by not taking advantage of the situation. Readers of research are rightly suspicious when data are omitted or transformed, as it is easy to imagine the temptation to trim and transform until the “right” results are obtained. The preceding guidelines should go a long way toward quelling such skepticism.

Uncontrollability

Thus far we have been concerned with the case in which a person is not aware of and does not intend to perceive or feel or behave in a certain way; it happens in the absence of a conscious intention. But what if the person was made aware of the effect? Could they control responses based on it if they wanted to? Uncontrollability of a process is another quality of automaticity, but one that need not follow from the others. In other words, it is very possible and probably even likely for one to be affected unintentionally by, say, the current environmental context (as in priming effects) but be able to counteract such effects on judgments or behavior if one becomes aware of the potential influence (Strack & Hannover, 1996). Devine (1989) showed that stereotype activation may be unintended, but with the appropriate values, motivation, and task, one

can control the effect of the stereotype on responses (see also Fiske, 1989; Moskowitz et al., 1999).

This leads to the general observation that although the initial activation events, such as in stereotyping, may not be easily if at all controlled, the overt responses based on those activated representations are controllable in most cases. Take the classic paradigm for studying uncontrollable activation: the Stroop color-word task (Stroop, 1935; see reviews in Logan, 1980; MacLeod, 1991). In this task, the participant is to name the color in which a word is presented. It is easily shown that people take longer when the word itself – which is irrelevant to the task of naming the color – is the name of a different color (e.g., the word RED presented in green ink).⁹ Researchers have shown that this effect holds for any stimuli to which the participant is perceptually sensitive, such as those related to his or her chronically accessible social constructs (Bargh & Pratto, 1986) or to discrepancies between his or her actual and ideal self-concepts (Higgins, VanHook, & Dorfman, 1988).

What is often overlooked in this paradigm is that the participants' actual responses in this task are overwhelmingly the correct ones. It is not that participants in the Stroop paradigm, instructed to name the color in which target words are presented, actually say (for example) “red” to the word RED in green ink; they do say “green” but just take longer to do so because of the need to inhibit the automatically activated competing response “red” (see Logan, 1980). So it has always been the case that findings of “uncontrollable” automatic effects refer not to uncontrollable responses but to uncontrollable internal activation events.

Again, the key is whether the individual is aware of the possibility of influence. If he or she is not, as in priming or stereotype-activation events, biased judgments and even behavior (Bargh et al., 1996) can be the result. But if the participant is made aware, he or she may be able to adjust for and control the effect (although overadjustment may occur; see Strack & Hannover, 1996). Take, for example, the classic study by Schwarz and Clore (1983) in which participants were contacted by telephone and asked questions about their life satisfaction. They were called either on a rainy or a sunny day, and if the interviewer did not mention the weather at all, it did affect their responses. Those contacted on a rainy day reported less satisfaction with their entire life than did those contacted on a sunny day, apparently misattributing their feelings due to the weather in the process. But if the experimenter casually referred to the current weather conditions, the effect disappeared. Calling the participants'

attention to the weather made it a piece of information in current working memory and more salient as a potential cause for their mood later on when they were asked about their life satisfaction.

An interesting variant on this theme is the *opposition* paradigm developed by Jacoby and his colleagues (e.g., Jacoby, 1991; Jacoby, Lindsay, & Toth, 1992). The essence of this procedure is to place conscious and unconscious influences in opposition to each other, so that the unconscious effects happen despite being contrary to intended, conscious purposes. In one study (Jacoby, Kelley, Brown, & Jasechko, 1989), for example, participants were exposed to a series of proper names as part of one experimental session. Half of the participants studied the list under full-attention conditions, whereas the remaining participants studied it under divided-attention conditions, having to perform a secondary task at the same time. The point of this attention manipulation was to decrease some participants' ability to remember later the names they had been shown.

Coming back to the lab the next day, the participants were asked to judge the fame of a list of names, which included new famous and new nonfamous names as well as some from the list of the previous day. Participants were told that all of the names they had studied the day before were nonfamous. Thus, if they consciously remembered seeing a name from that prior list, their response would be to say it was nonfamous. But participants from the divided-attention condition of the day before were less able to remember those names, and so less able to sort out whether the felt familiarity of those names came from their actual fame or from having seen them during the study phase of the experiment. As a result, the divided-attention condition participants were more likely than the full-attention participants to mistakenly say that the previous day's nonfamous names were actually famous – a demonstration, the authors concluded, of “becoming famous overnight.”

Note that neither the current weather conditions in the Schwarz and Clore study nor the original list of nonfamous names in the Jacoby *et al.* (1989) experiment were presented subliminally to participants, below their threshold of conscious awareness. All of the influential information was originally available to consciousness. The subliminality or supraliminality of the influential stimulus was not the critical factor in being influenced unintentionally and being unable to attempt to control that influence, but rather the participants' awareness of the potential effect of that (consciously perceived) stimulus.

How Control Attempts Can Produce Uncontrollability

Wegner and his colleagues (e.g., Ansfield & Wegner, 1996; Wegner, 1994; Wegner & Erber, 1992) generated a substantial body of evidence on uncontrollable processing effects. The basic experimental technique involves having participants engage in an attention-demanding secondary task while they are trying to prevent something from happening. Wegner's (1994) ironic process model makes the specific prediction that distraction and other strains on attentional capacity actually increase the likelihood that the counterintentional process will occur. That is, trying not to do something involves keeping in mind what it is that one does not want to happen, in order to maintain vigilance against it. But this has the ironic side effect of increasing the activation or accessibility of precisely those thoughts and behavior representations that one desires to control or prevent. Because the act of inhibiting or controlling them is effortful and attention demanding (Logan, 1980; Posner & Snyder, 1975; Shallice, 1972), trying not to do something under divided attention conditions will often have the ironic effect of making it more, not less, likely that one will do it. This is because one is left with the increased activation without the inhibition.

Ansfield and Wegner (1996) reported a series of experiments based on the Chevreul pendulum illusion, in which one is told to keep a pendulum still and not to let it move in a certain direction. As predicted by ironic process theory, having participants count backward from 1,000 by 7s while holding the pendulum caused the pendulum to move – as if by magic – exactly in the unintended direction. Ironic process theory identified a very large domain of uncontrollable mental processes, all of those one intends to control but cannot because of a current deficit in the attentional capacity needed to do so.

Conclusions

Priming is a very useful technique for studying the role played by situational context in cognition, motivation, and behavior. Such contextual effects are, if anything, more pervasive in everyday life than many social psychological theories allow. One's ongoing stream of consciousness continually creates ripples of influence that persist well after the conscious focus has flowed on to other things. And our conscious goals and purposes also continue to influence us after their originally intended task has been completed or abandoned.

Priming is also used to experimentally manipulate states of mind that are analogous to individual differences in automatic processing. One can select people based on these chronic differences, such as those high on achievement

motivation or those with a chronically accessible trait construct for honesty, and compare their performance on a task or their perceptions of a target person with those of participants without these chronic states. However, these groups of individuals could well differ in other ways as well, and they are self-selecting into the experimental conditions. A researcher's confidence in the focal independent variable as the real cause of an effect in individual difference research is bolstered if he or she can also produce the effect experimentally. Thus priming research is a natural complement to automaticity research.

The importance of studying automaticity resides in the ecological importance of the particular quality of automaticity that is under scrutiny. That is, it is important to study the efficiency or attention-free nature of a process when one wants to see if it would occur even in cognitively busy circumstances, and it is our feeling that these conditions are more the rule than the exception in life. And it is important to study whether a process occurs unintentionally because of the implications it has, in conjunction with lack of awareness, for the individual's ability to control it. If the process only happens when the person intends it, those with good intentions have nothing to fear. But in many cases good intentions go for naught because the person does not choose and is not aware of the perceptual or motivational process affecting him or her. And this lack of both intention and awareness may preclude controllability of the process.

Research into such automatic effects helped raise the general public's consciousness in the 1970s and 1980s about the possibility of nonconscious bias, especially in racial and gender stereotyping. Further study of these unseen hands of automatic influence can only continue to do such good. After all, it is only with such knowledge and awareness that one can hope to counteract those influences. An exciting contemporary trend in research, in fact, is aimed at discovering the conditions under which unwanted automatic influences, as in stereotyping, can be controlled or even changed.

But not all automatic influences are unwanted and counterproductive – quite the opposite. There is a natural tendency to assume, based on the findings of an automatic or nonconscious role in such social and personal problems as prejudice, sexual harassment, depression, and addictions, that automatic mental processes are always associated with negative outcomes, and conscious mental control with positive outcomes. Indeed, several influential authors have made just this argument (e.g., Bandura, 1986; Langer, 1989; Mischel, Cantor, & Feldman, 1996). Yet it is the natural purview of social psychologists to study social problems and for clinicians to study failures of self-regulation, and so the

problematic ones are likely to be overrepresented in the roll call of researched automatic phenomena in their research domains.

Habits of thought and behavior can be helpful as well as harmful: William James (1890) famously advised the young to make habitual as soon as possible all the useful behaviors one could. Just as negative stereotypes can be activated automatically, so too can chronic fairness motives (Moskowitz et al., 1999). Just as depressed people think about themselves automatically in negative terms, so too do nondepressed people think about themselves in automatically positive terms (Bargh & Tota, 1988), which turns out to be an important component of psychological health (e.g., Taylor & Brown, 1988). Therefore, another good tack for future research – besides the continued probe of how to control undesired automatic and contextual (priming) effects – might be to investigate the roles played by priming and automaticity in psychological health and socially constructive behavior. After all, nonconscious phenomena can be created and developed, as well as controlled and changed.

References

- Aarts, H., Custers, R., & Marien, H. (2008). Preparing and motivating behavior outside of awareness. *Science*, 319, 1639.
- Allport, F. H. (1955). *Theories of perception and the concept of structure*. New York: Wiley.
- Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, 89, 369–406.
- Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. New York: Winston.
- Ansfield, M. E., & Wegner, D. M. (1996). The feeling of doing. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action: Linking cognition and motivation to behavior* (pp. 482–506). New York: Guilford Press.
- Balota, D. A., & Lorch, Jr., R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12, 336–345.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

- Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 3–51). New York: Guilford Press.
- Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (2nd ed., Vol. 1, pp. 1–40). Hillsdale, NJ: Erlbaum.
- Bargh, J. A. (1996). Principles of automaticity. In E. T. Higgins & A. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 169–183). New York: Guilford Press.
- Bargh, J. A. (1997). The automaticity of everyday life. In R. S. Wyer, Jr. (Ed.), *Advances in social cognition* (Vol. 10, pp. 1–61). Mahwah, NJ: Erlbaum.
- Bargh, J. A. (2006). What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. *European Journal of Social Psychology*, 36, 147–168.
- Bargh, J. A., Bond, R. N., Lombard, W. J., & Tota, M. E. (1986). The additive nature of chronic and temporary sources of construct accessibility. *Journal of Personality and Social Psychology*, 50, 869–878.
- Bargh, J. A., Chaiken, S., Govender, R., & Pratto, F. (1992). The generality of the automatic attitude activation effect. *Journal of Personality and Social Psychology*, 62, 893–912.
- Bargh, J. A., Chaiken, S., Raymond, P., & Hymes, C. (1996). The automatic evaluation effect: Unconditionally automatic attitude activation with a pronunciation task. *Journal of Experimental Social Psychology*, 32, 185–210.
- Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, 71, 230–244.
- Bargh, J. A., & Ferguson, M. J. (2000). Beyond behaviorism: On the automaticity of higher mental processes. *Psychological Bulletin*, 126, 925–945.
- Bargh, J. A., & Gollwitzer, P. M. (1994). Environmental control of goal-directed action: Automatic and strategic contingencies between situations and behavior. In W. Spaulding (Ed.), *Integrations of motivation and cognition*:

The Nebraska Symposium on Motivation (Vol. 41, pp. 71–124). Lincoln: University of Nebraska Press.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A. Y., Barndollar, K., & Trötschel, R. (2001). Bypassing the will: Automatic and controlled self-regulation. *Journal of Personality and Social Psychology*, 81, 1014–1027.

Bargh, J. A., & Pietromonaco, P. (1982). Automatic information processing and social perception: The influence of trait information presented outside of conscious awareness on impression formation. *Journal of Personality and Social Psychology*, 43, 137–149.

Bargh, J. A., & Pratto, F. (1986). Individual construct accessibility and perceptual selection. *Journal of Experimental Social Psychology*, 22, 293–311.

Bargh, J. A., Raymond, P., Pryor, J., & Strack, F. (1995). The attractiveness of the underling: An automatic power–sex association and its consequences for sexual harassment and aggression. *Journal of Personality and Social Psychology*, 68, 768–781.

Bargh, J. A., & Thein, R. D. (1985). Individual construct accessibility, person memory, and the recall-judgment link: The case of information overload. *Journal of Personality and Social Psychology*, 49, 1129–1146.

Bargh, J. A., & Tota, M. E. (1988). Context-dependent automatic processing in depression: Accessibility of negative constructs with regard to self but not others. *Journal of Personality and Social Psychology*, 54, 925–939.

Bator, R. J., & Cialdini, R. B. (1995). *Priming a consistency motivation enhances cognitive dissonance effects*. Manuscript submitted for publication, Arizona State University.

Beck, A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects*. New York: Harper & Row.

Blair, I., & Banaji, M. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology*, 70, 1142–1163.

Boring, E. G. (1950). *A history of experimental psychology* (2nd ed.). New York: Appleton-Century-Crofts.

Bousfield, A. K., & Bousfield, W. A. (1966). Measurement of clustering and

- sequential constancies in repeated free recall. *Psychological Reports*, 19, 935–942.
- Bower, G. H., Lesgold, A. M., & Tieman, D. (1969). Grouping operations in free recall. *Journal of Verbal Learning and Verbal Behavior*, 8, 481–493.
- Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters*. New York: Wiley.
- Brenner, M. (1973). The next-in-line effect. *Journal of Verbal Learning and Verbal Behavior*, 12, 320–323.
- Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer, Jr. (Eds.), *Advances in social cognition* (Vol. 1, pp. 1–36). Hillsdale, NJ: Erlbaum.
- Broadbent, D. E. (1958). *Perception and communication*. London: Pergamon.
- Bruner, J. S. (1957). On perceptual readiness. *Psychological Review*, 64, 123–152.
- Bruner, J. S., & Postman, L. (1947). Value and need as organizing factors in perception. *Journal of Abnormal and Social Psychology*, 42, 33–44.
- Carver, C. S., Ganellen, R. J., Froming, W. J., & Chambers, W. (1983). Modeling: An analysis in terms of category accessibility. *Journal of Experimental Social Psychology*, 19, 103–121.
- Chaiken, S., & Bargh, J. A. (1993). Occurrence versus moderation of the automatic attitude activation effect: Reply to Fazio. *Journal of Personality and Social Psychology*, 64, 759–765.
- Chaiken, S., Giner-Sorolla, R., & Chen, S. (1996). Beyond accuracy: Defense and impression motives in heuristic and systematic information processing. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action* (pp. 553–578). New York: Guilford Press.
- Chartrand, T. L., & Bargh, J. A. (1996). Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of Personality and Social Psychology*, 71, 164–178.
- Chen, S., Shechter, D., & Chaiken, S. (1996). Getting at the truth or getting along: Accuracy and impression-motivated heuristic and systematic

- processing. *Journal of Personality and Social Psychology*, 71, 262–275.
- Cialdini, R. B. (1994, October). *The strain for consistency: A history, a measure, and a surprise*. Plenary address to the annual meetings of the Society for Experimental Social Psychology, Lake Tahoe, NV.
- Costin, F. (1969). The scrambled sentence test: A group measure of hostility. *Educational and Psychological Measurement*, 29, 461–468.
- Dalrymple-Alford, E. C. (1970). The measurement of clustering in free recall. *Psychological Bulletin*, 1, 32–34.
- Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, 70, 80–90.
- Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56, 680–690.
- Dijksterhuis, A., Chartrand, T. L., & Aarts, H. (2007). Effects of priming and perception on social behavior and goal pursuit. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 51–131). New York: Psychology Press.
- Dixon, N. F. (1971). *Subliminal perception: The nature of a controversy*. New York: McGraw-Hill.
- Dodge, K. A. (1993). Social-cognitive mechanisms in the development of conduct disorder and depression. *Annual Review of Psychology*, 44, 559–584.
- Doyen, S., Klein, O., Pichon, C-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, 7, e29081. doi:10.1371/journal.pone.0029081.
- Duckworth, K. L., Bargh, J. A., Garcia, M., & Chaiken, S. (2002). The automatic evaluation of novel stimuli. *Psychological Science*, 13, 513–519.
- Dunker, K. (1945). On problem solving. *Psychological Monographs*, 58 (5, Whole No. 270).
- Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. New York: Harcourt Brace Jovanovich.
- Edwards, K. (1990). The interplay of affect and cognition in attitude formation and change. *Journal of Personality and Social Psychology*, 59, 202–216.

- Erdelyi, M. H. (1974). A new look at the New Look: Perceptual defense and vigilance. *Psychological Review*, 81, 1–25.
- Erdley, C. A., & D'Agostino, P. R. (1988). Cognitive and affective components of automatic priming effects. *Journal of Personality and Social Psychology*, 54, 741–747.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87, 215–251.
- Fazio, R. H. (1990). A practical guide to the use of response latencies in social psychological research. In C. Hendrick & M. S. Clark (Eds.), *Review of personality and social psychology* (Vol. 11, pp. 74–97). Newbury Park, CA: Sage.
- Fazio, R. H. (1993). Variability in the likelihood of automatic attitude activation: Data reanalysis and commentary on Bargh, Chaiken, Govender, and Pratto (1992). *Journal of Personality and Social Psychology*, 64, 753–758.
- Fazio, R. H., Chen, J., McDonel, E. C., & Sherman, S. J. (1982). Attitude accessibility, attitude-behavior consistency, and the strength of the object-evaluation association. *Journal of Experimental Social Psychology*, 18, 339–357.
- Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, 69, 1013–1027.
- Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, 50, 229–238.
- Fischer, B., & Weber, H. (1993). Express saccades and visual attention. *Behavioral and Brain Sciences*, 16, 553–610.
- Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, 38, 889–906.
- Fiske, S. T. (1989). Examining the role of intent: Toward understanding its role in stereotyping and prejudice. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 253–283). New York: Guilford Press.

- Fiske, S. T., & Neuberg, S. E. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). San Diego, CA: Academic Press.
- Fowler, C. A., Wolford, G., Slade, R., & Tassinary, L. (1981). Lexical access with and without awareness. *Journal of Experimental Psychology: General*, 110, 341–362.
- Gilbert, D. T. (1989). Thinking lightly about others: Automatic components of the social inference process. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 189–211). New York: Guilford Press.
- Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, 60, 509–517.
- Gilbert, D. T., & Osborne, R. E. (1989). Thinking backward: Some curable and incurable consequences of cognitive busyness. *Journal of Personality and Social Psychology*, 57, 940–949.
- Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When persons perceivers meet persons perceived. *Journal of Personality and Social Psychology*, 54, 733–740.
- Gollwitzer, P. M. (1990). Action phases and mindsets. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition* (Vol. 2, pp. 53–92). New York: Guilford Press.
- Gollwitzer, P. M., & Bargh, J. A. (Eds.). (1996). *The psychology of action*. New York: Guilford Press.
- Gollwitzer, P. M., Heckhausen, H., & Steller, B. (1990). Deliberative and implemental mindsets: Cognitive tuning toward congruous thoughts and information. *Journal of Personality and Social Psychology*, 59, 1119–1127.
- Gollwitzer, P. M., & Moskowitz, G. B. (1996). Goal effects on action and cognition. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 361–399). New York: Guilford Press.
- Grand, S., & Segal, S. J. (1966). Recovery in the absence of recall. *Journal of Experimental Psychology*, 72, 138–144.

- Greenwald, A. G. (1992). New Look 3: Unconscious cognition reclaimed. *American Psychologist*, 47, 766–779.
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102, 4–27.
- Greenwald, A. G., Draine, S. C., & Abrams, R. L. (1996). Three cognitive markers of unconscious semantic activation. *Science*, 273, 1699–1702.
- Greenwald, A. G., Klinger, M. R., & Liu, T. J. (1989). Unconscious processing of dichoptically masked words. *Memory and Cognition*, 17, 35–47.
- Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin*, 75, 124–129.
- Hamilton, D. L., Katz, L. B., & Leirer, V. O. (1980a). Cognitive representation of personality impression: Organizational processes in first impression formation. *Journal of Personality and Social Psychology*, 39, 1050–1063.
- Hamilton, D. L., Katz, L. B., & Leirer, V. O. (1980b). Organizational processes in impression formation. In R. Hastie, T. M. Ostrom, E. B. Ebbesen, R. S. Wyer, Jr., D. L. Hamilton, & D. E. Carlston (Eds.), *Person memory: The cognitive basis of social perception* (pp. 121–153). Hillsdale, NJ: Erlbaum.
- Hastie, R., & Kumar, P. A. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality and Social Psychology*, 37, 25–38.
- Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review*, 93, 258–268.
- Herr, P. M. (1986). Consequences of priming: Judgment and behavior. *Journal of Personality and Social Psychology*, 51, 1106–1115.
- Herr, P. M., Sherman, S. J., & Fazio, R. H. (1984). On the consequences of priming: Assimilation and contrast effects. *Journal of Experimental Social Psychology*, 19, 323–340.
- Higgins, E. T. (1989). Knowledge accessibility and activation: Subjectivity and suffering from unconscious sources. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 75–123). New York: Guilford Press.
- Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and

- salience. In E. T. Higgins & A. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). New York: Guilford Press.
- Higgins, E.T., Bargh, J. A., & Lombardi, W. (1985). Nature of priming effects on categorization. *Journal of Experimental Social Psychology*, 11, 59–69.
- Higgins, E. T., & Chaires, W. M. (1980). Accessibility of inter-relational constructs: Implications for stimulus encoding and creativity. *Journal of Experimental Social Psychology*, 16, 348–361.
- Higgins, E. T., & King, G. A. (1981). Accessibility of social constructs: Information-processing consequences of individual and contextual variability. In N. Cantor & J. F. Kihlstrom (Eds.), *Personality, cognition, and social interaction* (pp. 69–122). Hillsdale, NJ: Erlbaum.
- Higgins, E. T., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, 13, 141–154.
- Higgins, E. T., VanHook, E., & Dorfman, D. (1988). Do self-attributes form a cognitive structure? *Social Cognition*, 6, 177–217.
- Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey. *Behavioral and Brain Sciences*, 9, 1–66.
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30, 513–541.
- Jacoby, L. L., Kelley, C., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology*, 56, 326–338.
- Jacoby, L. L., Lindsay, D. S., & Toth, J. P. (1992). Unconscious influences revealed: Attention, awareness, and control. *American Psychologist*, 47, 802–809.
- James, W. (1890). *Principles of psychology*. New York: Holt.
- Johnson, M. K., & Hasher, L. (1987). Human learning and memory. *Annual Review of Psychology*, 38, 631–668.
- Kantowitz, B. H. (1974). Double stimulation. In B. H. Kantowitz (Ed.), *Human*

- information processing* (pp. 320–342). Hillsdale, NJ: Erlbaum.
- Koestler, A. (1967). *The ghost in the machine*. London: Hutchinson & Co.
- Koffka, K. (1922). Perception: An introduction to the Gestalt-theorie. *Psychological Bulletin*, 19, 531–585.
- Koriat, A., & Feuerstein, N. (1976). The recovery of incidentally acquired information. *Acta Psychologica*, 40, 463–464.
- Langer, E. J. (1978). Rethinking the role of thought in social interaction. In J. H. Harvey, W. I. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 2, pp. 35–58). Hillsdale, NJ: Erlbaum.
- Langer, E. J. (1989). *Mindfulness*. New York: Allyn & Bacon.
- Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior: The Hixon symposium* (pp. 112–136). New York: Wiley.
- Linville, P. (1996). Attention inhibition: Does it underlie ruminative thought? In R. S. Wyer, Jr. (Ed.), *Advances in social cognition* (Vol. 9, pp. 121–133). Mahwah, NJ: Erlbaum.
- Logan, G. D. (1980). Attention and automaticity in Stroop and priming tasks: Theory and data. *Cognitive Psychology*, 12, 523–553.
- Logan, G. D., & Cowan, W. (1984). On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review*, 91, 295–327.
- Lombardi, W. J., Higgins, E. T., & Bargh, J. A. (1987). The role of consciousness in priming effects on categorization. *Personality and Social Psychology Bulletin*, 13, 411–429.
- MacLeod, C. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109, 163–203.
- Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology*, 23, 77–87.
- Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, 66, 37–17.

- Marcel, A. J. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology*, 15, 197–237.
- Marks, M. M., Sinclair, R. C., & Wellens, T. R. (1991). The effect of completing the Beck Depression Inventory on self-reported mood state: Contrast and assimilation. *Personality and Social Psychology Bulletin*, 17, 457–465.
- Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology*, 35, 63–78.
- Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90, 227–234.
- Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, 80, 252–283.
- Mischel, W., Cantor, N., & Feldman, S. (1996). Principles of self-regulation: The nature of willpower and self-control. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 329–360). New York: Guilford Press.
- Moskowitz, G. B., Gollwitzer, P. M., Wasel, W., & Schaal, B. (1999). Preconscious Control of Stereotype activation through chronic egalitarian goals. *Journal of Personality and Social Psychology*, 77, 167–184.
- Murphy, M. D. (1979). Measurement of category clustering in free recall. In C. R. Puff (Ed.), *Memory organization and structure* (pp. 51–83). San Diego, CA: Academic Press.
- Murphy, S. T., Monahan, J. L., & Zajonc, R. B. (1995). Additivity of nonconscious affect: Combined effects of priming and exposure. *Journal of Personality and Social Psychology*, 69, 589–602.
- Murphy, S. T., & Zajonc, R. B. (1993). Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology*, 64, 723–739.
- Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, 106, 226–254.
- Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

- Neuberg, S. L. (1988). Behavioral implications of information presented outside of conscious awareness: The effect of subliminal presentation of trait information on behavior in the Prisoner's Dilemma Game. *Social Cognition*, 6, 207–230.
- Neumann, O. (1984). Automatic processing: A review of recent findings and a plea for an old theory. In W. Prinz & A. F. Sanders (Eds.), *Cognition and motor processes* (pp. 255–293). Berlin: Springer-Verlag.
- Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.
- Newman, L. S., & Uleman, J. S. (1990). Assimilation and contrast effects in spontaneous trait inference. *Personality and Social Psychology Bulletin*, 16, 224–240.
- Niedenthal, P. M. (1990). Implicit perception of affective information. *Journal of Experimental Social Psychology*, 26, 505–527.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.
- Norman, D. A. (1968). Toward a theory of memory and attention. *Psychological Review*, 75, 522–536.
- Ostrom, T. M., Pryor, J. B., & Simpson, D.D. (1981). The organization of social information. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), *Social cognition: The Ontario Symposium* (Vol. 1, pp. 3–38). Hillsdale, NJ: Erlbaum.
- Over, H., & Carpenter, M. (2009). Eighteen-month-old infants show increased helping following priming with affiliation. *Psychological Science*, 20, 1189–1193.
- Perdue, C. W., Dovidio, J. F., Gurtman, M. B., & Tyler, R. B. (1990). Us and them: Social categorization and the process of intergroup bias. *Journal of Personality and Social Psychology*, 59, 475–486.
- Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 55–85). Hillsdale, NJ: Erlbaum.
- Postman, L., Bruner, J. S., & McGinnies, E. (1948). Personal values as selective

- factors in perception. *Journal of Abnormal and social Psychology*, 43, 142–154.
- Pratto, P., & Bargh, J. A. (1991). Stereotyping based on apparently individuating information: Trait and global components of sex stereotypes under attention overload. *Journal of Experimental Social Psychology*, 27, 26–47.
- Prinz, W. (1990). A common coding approach to perception and action. In O. Neumann & W. Prinz (Eds.), *Relationships between perception and action* (pp. 167–201). Heidelberg, Germany: Springer-Verlag.
- Rayner, K. (1978). Foveal and parafoveal cues in reading. In J. Requin (Ed.), *Attention and performance VIII* (pp. 149–161). Hillsdale, NJ: Erlbaum.
- Rizzolatti, G., & Sinigaglia, C. (2008). *Mirrors in the brain: How our minds share actions and emotions* (F. Anderson, Tr.). New York: Oxford University Press.
- Robertson, C. (1995). Modeling recall: Clustering and order effects. *British Journal of Mathematical and Statistical Psychology*, 48, 29–50.
- Roenker, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin*, 1, 45–18.
- Rogers, T. B. (1974). An analysis of two central stages underlying responding to personality items: The self-referent decision and response selection. *Journal of Research in Personality*, 8, 128–138.
- Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.
- Roskos-Ewoldsen, D. R., & Fazio, R. H. (1992). On the orienting value of attitudes: Attitude accessibility as a determinant of an object's attraction of visual attention. *Journal of Personality and Social Psychology*, 63, 198–211.
- Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. New York: McGraw-Hill.
- Rothbart, M. (1981). Memory processes and social beliefs. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 272–298). Hillsdale, NJ: Erlbaum.
- Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of*

- Experimental Psychology: Learning, Memory, and Cognition*, 13, 501–518.
- Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 45, 513–523.
- Seamon, J. G., Brody, N., & Kauff, D. M. (1983). Affective discrimination of stimuli that are not recognized: Effects of shadowing, masking, and cerebral laterality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9, 544–555.
- Segal, S. J. (1967). The priming of association test responses. *Journal of Verbal Learning and Verbal Behavior*, 6, 216–221.
- Segal, S. J., & Gofer, C. N. (1960). The effect of recency and recall on word association. *American Psychologist*, 15, 451.
- Shallice, T. (1972). Dual functions of consciousness. *Psychological Review*, 79, 383–393.
- Shevrin, H. (1992). Subliminal perception, memory, and consciousness: Cognitive and dynamic perspectives. In R. Bornstein & T. Pittman (Eds.), *Perception without awareness* (pp. 123–142). New York: Guilford Press.
- Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84, 127–190.
- Skelton, J. A., & Strohmets, D. B. (1990). Priming symptom reports with health-related cognitive activity. *Personality and Social Psychology Bulletin*, 16, 449–464.
- Smith, E. R. (1994). Procedural knowledge and processing strategies in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (2nd ed., Vol. 1, pp. 99–152). Hillsdale, NJ: Erlbaum.
- Smith, E. R., Branscombe, N., & Bormann, C. (1988). Generality of the effects of practice on social judgment tasks. *Journal of Personality and Social Psychology*, 54, 385–395.
- Smith, E. R., & Lerner, M. (1986). Development of automatism of social judgments. *Journal of Personality and Social Psychology*, 50, 246–259.
- Smith, E. R., Stewart, T. L., & Buttram, R. T. (1992). Inferring a trait from a

- behavior has long-term, highly specific effects. *Journal of Personality and Social Psychology*, 62, 753–759.
- Sorrentino, R. M., & Higgins, E. T. (1986). Motivation and cognition: Warming up to synergism. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition* (Vol. 1, pp. 3–19). New York: Guilford Press.
- Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs*, 74, Whole No. 498.
- Srull, T. K. (1981). Person memory: Some tests of associative storage and retrieval models. *Journal of Experimental Psychology: Human Learning and Memory*, 7, 140–162.
- Srull, T. K. (1984). Methodological techniques of the study of person memory and social cognition. In R. S. Wyer, Jr. & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 73–150). Hillsdale, NJ: Erlbaum.
- Srull, T. K., Lichtenstein, M., & Rothbart, M. (1985). Associated storage and retrieval processes in person memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 316–345.
- Srull, T. K., & Wyer, Jr., R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology*, 37, 1660–1672.
- Srull, T. K., & Wyer, Jr., R. S. (1980). Category accessibility and social perception: Some implications for the study of person memory and interpersonal judgments. *Journal of Personality and Social Psychology*, 38, 841–856.
- Srull, T. K., & Wyer, Jr., R. S. (1989). Person memory and judgment. *Psychological Review*, 96, 58–83.
- Storms, L. H. (1958). Apparent backward association: A situational effect. *Journal of Experimental Psychology*, 55, 390–395.
- Strack, R., & Hannover, B. (1996). Awareness of influence as a precondition for implementing correctional goals. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action* (pp. 579–596). New York: Guilford Press.
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643–662.

- Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, 103, 193–210.
- Tice, D. M., Butler, J. L., Muraven, M. B., & Stillwell, A. M. (1995). When modesty prevails: Differential favorability of self-presentation to friends and strangers. *Journal of Personality and Social Psychology*, 69, 1120–1138.
- Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, 12, 242–248.
- Trope, T. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review*, 93, 239–257.
- Turvey, M. T. (1973). On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. *Psychological Review*, 80, 1–52.
- Uleman, J. S., Hon, A., Roman, R. J., & Moskowitz, G. B. (1996). On-line evidence for spontaneous trait inferences at encoding. *Personality and Social Psychology Bulletin*, 22, 377–394.
- Wegener, D. T., & Petty, R. E. (1995). Flexible correction processes in social judgment: The role of naive theories in corrections for perceived bias. *Journal of Personality and Social Psychology*, 68, 36–51.
- Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review*, 101, 34–52.
- Wegner, D. M., & Bargh, J. A. (1998). Control and automaticity in social life. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., pp. 446–496). Boston: McGraw-Hill.
- Wegner, D. M., & Erber, R. (1992). The hyperaccessibility of suppressed thoughts. *Journal of Personality and Social Psychology*, 63, 903–912.
- Wicklund, R. A., & Brehm, J. W. (1976). *Perspectives on cognitive dissonance*. Hillsdale, NJ: Erlbaum.
- Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, 116, 117–142.
- Wilson, T. D., & Capitman, J. A. (1982). Effects of script availability on social

behavior. *Personality and Social Psychology Bulletin*, 8, 11–19.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Wyer, Jr., R. S., & Carlston, D. E. (1979). *Social cognition, inference, and attribution*. Hillsdale, NJ: Erlbaum.

Wyer, Jr., R. S., & Gordon, S. E. (1982). The recall of information about persons and groups. *Journal of Experimental Social Psychology*, 18, 128–164.

Wyer, Jr., R. S., & Srull, T. K. (1989). *Memory and cognition in its social context*. Hillsdale, NJ: Erlbaum.

* This chapter was written while the senior author was a visiting professor at the University of Konstanz. Its preparation was supported in part by Grant SBR-9409448 from the National Science Foundation and by a fellowship from the Alexander von Humboldt Foundation. We thank the editors and Miguel Brendl, Peter Gollwitzer, Asher Koriath, Katelyn McKenna, James Uleman, Wolfgang Wasel, and Gifford Weary for their advice and suggestions. Correspondence should be directed to John A. Bargh at Department of Psychology, Yale University, or Tanya L. Chartrand, Fuqua School of Business, Duke University.

¹ In fact, Higgins and Chaires (1980) demonstrated how solutions to the Duncker candle problem could be produced using the more modern priming techniques discussed in this chapter. A participant repeatedly exposed to the word “or” as part of an apparently unrelated experiment was more likely to see a box of tacks as two separate objects – a box and some tacks – compared with participants previously exposed to the word “and.” This was shown by the or-primed participants’ greater success rate in solving a puzzle in which the box had to be tacked to a wall in order to form a platform for the candle.

² Note that this is not normally a problem in subliminal priming, in which the same set of words directly related to the primed concept can be repeatedly presented over the course of the priming task (see Bargh & Pietromonaco, 1982).

³ Perdue, Dovidio, Gurtman, and Tyler (1990) presented words (related to the

group concepts of “us” and “them”) foveally for 55 ms, with immediate pattern-masking, but as they did not run awareness checks (basing claims of subliminality instead on the reports of pretest participants), one should be extremely cautious in the use of such a lengthy foveal presentation time. Seamon, Brody, and Kauff (1983), for example, found greater than chance recognition of polygons presented foveally (and pattern-masked) at 5 ms.

⁴ When using CRT monitors for subliminal priming, the minimum presentation time is constrained by the monitor hardware, specifically the screen refresh rate. For example, a 60 Hz CRT monitor updates its display 60 times a second, or once every 16.7 ms; a 70 Hz monitor every 1/70th of a second or 14.3 ms. Even if the program controlling the display instructs that a stimulus be displayed for a shorter time than this, the stimulus will nonetheless be displayed for the full duration of the screen refresh cycle. Note this is not an issue for LCD screens or laptops.

⁵ In fact, if participants are aware of the source or reason for activation when performing the second task, and believe it has an unwanted influence, they may correct or overcorrect for that influence.

⁶ There is a difference of opinion about the generality of the effect and whether it is moderated by the “strength” of the attitude in memory (see, e.g., Chaiken & Bargh, 1993; Fazio, 1993), but a great deal of consensus as to the existence of the automaticity effect itself.

⁷ It might be questioned why it is acceptable to routinely perform such trimming with reaction time data when one is not routinely permitted to trim outliers in other forms of dependent measures (e.g., responses at the opposite end of questionnaire scales as most other responses). Although we do not claim to offer a definitive answer to this provocative objection, in a first pass at an answer we would point to the usually small, though meaningful, differences between conditions typically obtained with reaction time methods, which can be easily swamped and distorted by just a single outlier; secondly, unlike outliers in questionnaires that are the product of conscious choice, those in reaction time studies are most usually errors of some form and not psychologically meaningful (e.g., a response time of 4 s to pronounce “elephant” for a native English-

speaker).

⁸ For an example of careful outlier analysis and elimination, see Uleman, Hon, Roman, and Moskowitz (1996, pp. 381–382). These researchers also provide useful guidelines for dropping participants with high error rates and for eliminating the effects of practice, fatigue, and boredom that can occur during experiments with many response trials.

⁹ There is no reason to limit this technique to the task of color-naming; the logic applies equally well to any task in which an irrelevant dimension of the stimuli suggests the same or a competing response to that dictated by the relevant dimension. In the original Stroop task, the meaning of the stimulus word is an irrelevant dimension, and its color the relevant dimension, but participants cannot help but process the irrelevant feature. But if the task is instead to indicate whether a stimulus word was presented above or below a fixation point, then the word “above” facilitates response times (compared with other words) when presented above the fixation point and slows down response times when presented below the fixation point (contrariwise for the word “below”; see Logan, 1980).

Appendix A: Examples of Scrambled Sentence Tests

Instructions: For each set of words below, make a grammatical four word sentence and write it down in the space provided.

For example:

flew eagle the plane around The eagle flew around.

(from Bargh, Chen, & Burrows, 1996, Experiment 2)

1. him was *worried* she always
2. from are *Florida* oranges temperature
3. ball the throw toss silently
4. shoes give replace *old* the
5. he observes occasionally people watches
6. be will swear *lonely* they
7. sky the seamless *grey* is

8. ate she it selfishly all
9. be to back *careful* better
10. prepare the gift wrap neatly
11. sew *sentimental* buy item the
12. he *wise* drops only seems
13. are we *stubborn* courteous sometimes
14. the push wash frequently clothes
15. us *bingo* sing play let
16. should now withdraw *forgetful* we
17. somewhat prepared I was retired
18. sunlight makes temperature *wrinkle* raisins
19. is *rigid* he usually studying
20. a have *traditional* wedding holiday
21. picked throw apples hardly the
22. drink this looks seems *bitter*
23. they obedient him often meet
24. there are they *conservative* going
25. knits *dependent* he occasionally them
26. studies she texts *ancient* him
27. helpless it hides there over
28. is he *gullible* plant so
29. *cautious* alone very are they
30. send I mail it over

(from Chartrand & Bargh, 1996, Experiment 1)

1. from are Florida *preserve* they
2. a smile parrot what great
3. watches *recalls* he occasionally people
4. ball the hoop toss normally
5. saw hammer he train the
6. good dislikes *recognizes* she deals
7. maintain she to composure try
8. should withdraw *keep* now we
9. the machine wash frequently clothes
10. somewhat *memory* prepared I was
11. save does *study* usually he
12. be to *remember* back careful
13. sky the seamless red is

14. a have June holiday wedding
15. they *retain* him often meet

Note: Words in italics are the critical priming stimuli (for the “elderly” stereotype and the goal of memorization, respectively); they are not italicized in the actual task.

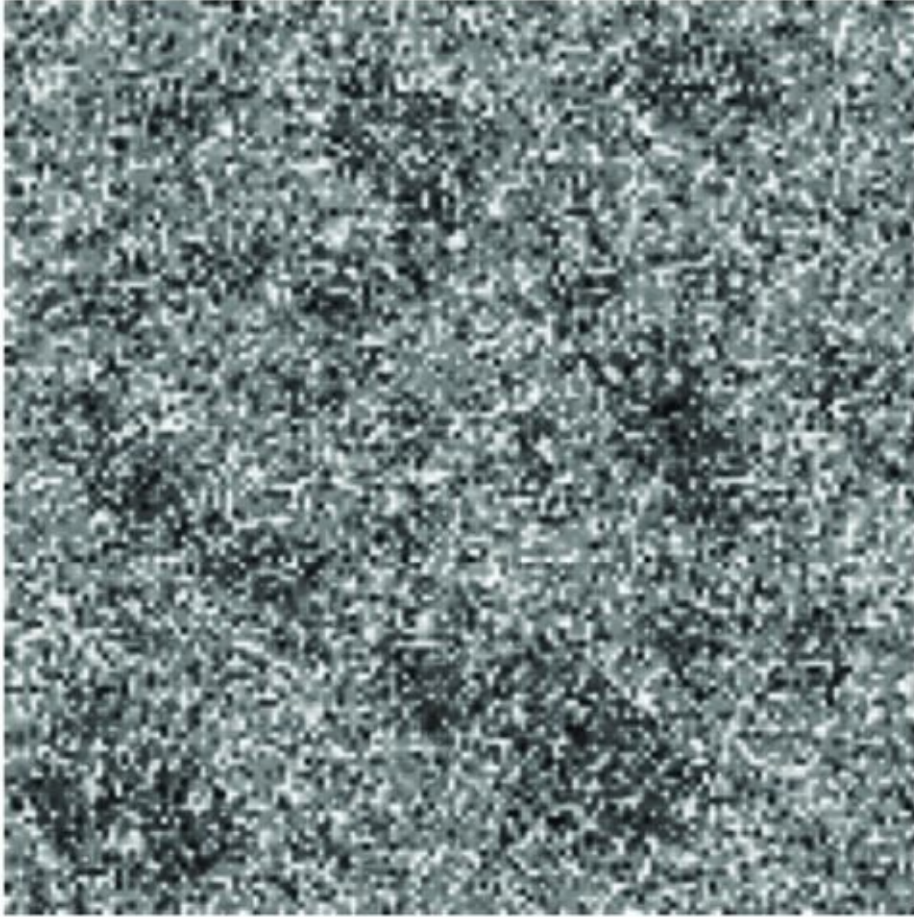
Appendix B: Example of Funneled Debriefing Procedure for Supraliminal Priming Task

The experimenter proceeds to ask the participant the following questions and records the answers given:

1. What do you think the purpose of this experiment was?
2. What do you think this experiment was trying to study?
3. Did you think that any of the tasks you did were related in any way?(if “yes”) In what way were they related?
4. Did anything you did on one task affect what you did on any other task? (if “yes”) How exactly did it affect you?
5. When you were completing the scrambled sentence test, did you notice anything unusual about the words?
6. Did you notice any particular pattern or theme to the words that were included in the scrambled sentence test?
7. What were you trying to do while reading the behavioral phrases on the computer monitor? Did you have any particular goal or strategy?

Source: Chartrand & Bargh ([1996](#), Experiment 1).

Appendix C: Example of a Subliminally Presented Pattern Used as a Forward and Backward Mask



Source: Fitzsimons, Chartrand, & Fitzsimons (2008).

Chapter fourteen Behavioral Observation and Coding

Richard E. Heyman, Michael F. Lorber, J. Mark Eddy and Tessa V. West

Behavioral Observation and Coding

Kurt Lewin (1951, p. 169) wrote that there is “nothing so practical as a good theory.” One could add that there is nothing so practical as a good theory-testing tool. We devote this chapter to one such tool – behavioral observation – that excels at both the identification of behaviors worth theorizing about and the testing of theories of behavior. This chapter provides an overview of behavioral observation, including the contexts researchers use when observing, the forms in which they record behaviors for analysis (e.g., coding), the methods available to document that different observers coded behaviors similarly (i.e., interrater agreement, an element of reliability), the necessity of establishing other forms of reliability as well as validity, and methods of analyzing behavioral observation data.

What Is Behavioral Observation?

The observation of behavior is at the center of all scientific inquiry in social psychology. Although there are a wide variety of methods that researchers use when observing, the term “behavioral observation” generally refers to a researcher seeing and/or hearing, and then systematically recording, the behaviors of an individual or group of individuals within a particular social context of interest, such as the classroom, the playground, the peer group, the home, the clinic, or the workplace. Typically, individuals are observed for relatively brief periods of time, but often on multiple occasions.

Sometimes observations are conducted “live.” More often, an audio or video recording is made (and sometimes transcribed into written form as well); observations are then conducted using one or more of these at the convenience of the researcher. During an observation, a researcher periodically summarizes

the physical and/or verbal behaviors of the participants of interest into specific categories using a clearly defined system of “codes” that are assigned based on a set of rules. Each code is used to mark the occurrence of a specific behavior or set of behaviors and, depending on the data collection technique, may be recorded in parallel with other information about the code (e.g., affective quality, start and stop times). The final result is a sequential record of the behaviors of one or more individuals.

Why Use This Research Method?

Some value behavioral observation data simply because it provides objective information about the frequencies of particular behaviors engaged in by a given individual. This might be important, for example, to a researcher interested in examining whether an intervention changed the frequency of certain targeted behaviors (e.g., factors that promote or inhibit bystander intervention; Penner, Dovidio, Piliavin, & Schroeder, 2005). Others value observation as a means to examine the relations between and among behaviors, either within individuals or among dyads or groups. This might be important, for example, to a researcher who is interested in whether the same behaviors are expressed during inter-and intraracial interactions, and whether perceivers apply the same meaning to these behaviors as a function of the racial composition of the interacting dyad (e.g., Gray, Mendes, & Denny-Brown, 2008). Thus, behavioral observations are useful for answering questions not only about individual *outcomes* but also about social interactional *processes*.

In many studies, outcomes and processes are measured through self-reports from one informant. The generalizability of findings from this measurement strategy may be limited. In an attempt to rectify this problem, a high value has been placed in recent years on the use of multiple informants and multiple assessment methods to measure psychological constructs of importance (e.g., Smith & Harris, 2006; see also Brewer & Crano, Chapter 2 in this volume). This approach is thought to lead to a more reliable and valid index of the “true score” of a construct. Introducing multiple informants into assessment batteries is relatively easy – various forms of self-report questionnaires on behaviors or behavioral patterns are readily available, or can be created relatively easily, for different reporters (e.g., parents, teachers, youth).

Although behavioral observation is a quite appealing method for some researchers, it does have its downsides. Even if an existing coding system is

identified for a new study, purchasing the necessary equipment, securing private coding space, and assembling and training a team of observers (i.e., a “coding team”) can be time consuming and expensive. Once a team is ready, collecting data in vivo, or collecting and storing video or audio records and transcribing those records, and then managing and analyzing the resulting data can also be quite costly.

Furthermore, although the focus of a typical coding team is usually on obtaining and maintaining interrater agreement (i.e., independent observers applying the same codes to a given stream of behavior), this is no guarantee that behavioral observation will generate reliable (i.e., stable) or valid (i.e., “true” measures) scores of constructs of interest in a given sample. Indicators derived from behavioral observation often are weakly correlated with self-report measures of the same constructs, and the meaning of this may be unclear. Finally, the existence of audio or video records creates ongoing human subjects issues related to the protection of confidentiality and anonymity. In short, despite their appeal, “observational data, compared with other forms of data, are unwieldy and messy” (Margolin, Oliver, Gordis, O’Hearn, Medina, Ghosh, & Morland, 1998, p. 29). Nevertheless, behavioral observation has been employed frequently over the past 50 years, particularly among psychologists interested in interpersonal and intergroup relations, human development, and close relationships .

Observational Settings

Observational settings exist along a continuum of researcher influence ranging from unfettered natural environments to tightly controlled experimental situations. Purely naturalistic situations have the advantage of being high in ecological validity (see Cialdini & Paluck, [Chapter 5](#) in this volume). Although researchers observing behavior in its natural environment still need to establish the reliability of their observations (e.g., consistency across observers, episodes, or settings), the real-world generalizability of such observations is self-evident. The more the researcher intervenes in the setting to be observed, the more has to be done to demonstrate that the setting produces externally valid results.

In the sections that follow we provide an overview of different degrees of researcher interventions into settings. As with any research tool, the validity of behavioral observation is situation dependent and can only be inferred from that tested, narrow use; it is not “proven” for all time (e.g., Haynes & O’Brien,

2000). Thus, behavioral observation cannot be said to be a valid assessment approach any more than questionnaires can be said to be a valid assessment approach.

Naturalistic Observation

Naturalistic observation has a long history in the study of animal (e.g., Lorenz, 1970, 1971) and human (e.g., Mead, 1928) behavior. Some researchers who favor this type of observation use a qualitative approach, where the coding system is not predetermined. Others use a quantitative approach, marked by the use of preset codes and precisely defined rules for their assignment.

One of the most important studies in social psychology – Festinger, Riecken, and Schacter's (1956) *When Prophecy Fails*, which focused on social interactions within a doomsday cult and proposed cognitive dissonance theory – used naturalistic observation. Observers ultimately were not outsiders, but rather became members of the social group being observed. There were no predetermined codes to classify behaviors. The observers who infiltrated the cult received only an overview of the study's purpose and the phenomena of highest interest. As Festinger *et al.* (1956, p. 248) described, “Problems of rigor and systematization in observation took a back seat in the hurly-burly of simply trying to keep up with a movement that often seemed to us to be ruled by whimsy.” The researchers also noted a common problem with many naturalistic studies: “[Observing] was frequently irritating because of the irrelevancies...that occupied vast quantities of time...[and] the repetitiousness of much that was said.” Observing surreptitiously without modern hidden recorders also necessitated observers taking frequent bathroom breaks or walks outdoors to absent themselves from the group to take notes.

Whereas Festinger and colleagues used naturalistic observation to examine an extraordinary social situation, most investigators use this approach to examine the ordinary (i.e., how interactions during normal life are related to particular outcomes of interest). To facilitate the collection of this type of information, Mehl and colleagues (e.g., Mehl, Pennebaker, Crow, Dabbs, & Price, 2001) developed a behavior observational paradigm that employs the Electronically Activated Recorder (EAR; Mehl, 2007; Mehl & Robbins, 2012). The EAR is an audio recorder that is worn in everyday settings and is programmed to make 30-second audio samples every 12.5 minutes (i.e., 5 minutes of recordings per hour).

The EAR has been used to examine questions such as “Do women really talk

more than men?” (From the research done so far, it appears that they do not; e.g., Mehl et al., 2007). The coding system developed for this paradigm, the Social Environment Coding of Sound Inventory (SECSI; Mehl & Pennebaker, 2003; Mehl et al., 2006), comprises four categories (with codes within those categories): (1) current location (e.g., home, outside), (2) activity (e.g., watching TV, eating), (3) social interaction (e.g., alone, talking on phone, in group), and (4) behavioral indicators of mood (e.g., laughing, crying, arguing). Although this work has produced important findings related to health, Mehl (2007, p. 370), drew the same conclusion as Festinger on the banality of observing life naturalistically: “One of my first ‘aha!’ experiences when we started doing EAR research was how ordinary and mundane real life really is. The sound files we obtained from participants first and foremost documented that for most people most of real life is not thrilling, glittery, and extraordinary.”

Another recent use of naturalistic observation was of families of dual-earning parents in California (Campos, Graesch, Repetti, Bradbury, & Ochs, 2009). Because the investigators had an overwhelming 35 hours of video from two weekdays per family, Campos *et al.* (2009) focused only on the two minutes captured when the partners reunited after their workdays and coded these simply (i.e., positive, negative, ignoring/distracted, reporting information, checking in about logistics). The authors also presented data from the “scan sampling” of family interactions, in which, every 10 minutes, observers noted the location of each family member. They found that working couples spend almost no time together without children. In later analyses, they found that men's, not women's, “neuroticism” (i.e., temperamental negativity) moderated the relationship between job stress and at-home behavior (Wang, Repetti, & Campos, 2011). For instance, men high in job stress but low in neuroticism were more socially withdrawn during their first hour home, but their interactions with their children were more intense .

Quasi-Naturalistic Observation

As implied by the Festinger and Mehl quotes, naturalistic observation often requires so much time that it is inefficient and impractical. Thus, observation typically occurs in situations that are not completely natural and uninfluenced by the investigator. When investigators use quasi-naturalistic observations, the generalizability of behavior is of the highest concern, and investigators attempt to influence the situation as little as possible.

The work of the Oregon Social Learning Center (OSLC) research team (e.g.,

Reid, Patterson, & Snyder, 2002) is a model of the development and refinement of a quasi-naturalistic observational paradigm. Starting in the late 1960s, OSLC researchers wanted to conduct naturalistic behavior observations of families but quickly learned that the natural world was not conducive to cost-effective data collection (Patterson, 1982). Family members typically disappeared or sat transfixed in front of a video screen when observers arrived (and this was usually a solitary television screen, long before the advent of other screen-related distractions in the home, such as smart phones, iPads, computers, video games, etc.). Out of necessity, eight rules (see Table 14.1) were imposed on families during their in-home observation sessions. Patterson (1982) noted that the rules transformed the otherwise typical environment into something close, but not identical, to the real world (i.e., those being observed were unnaturally constrained but otherwise acting naturally in their natural environment). This increases the quality of the data collected by increasing interaction but reduces generalizability slightly – exactly the kind of trade-off all researchers must weigh in designing protocols.

Table 14.1. Rules for Quasi-Naturalistic Family Observation Sessions

-
1. Everyone in the family must be present.
 2. No guests.
 3. The family is limited to two rooms.
 4. The observers will wait only 10 minutes for all to be present in the two rooms.
 5. Telephone: No calls out; briefly answer incoming calls.
 6. No TV.
 7. No talking to observers while they are coding.
 8. Do not discuss anything with the observers that relates to your problems or the procedures you are using to deal with them.
-

Source: Reid (1978, p. 8).

OSLC developed its quasi-naturalistic paradigm through trial and error, guided by both the empirical literature and their theoretical model. The researchers were most interested in children's aversive and aggressive behaviors and their parents' responses to these behaviors. To increase the chance of observing such interactions, dinnertime was chosen as the setting to observe, because earlier studies had found that mothers reported the most conflict with their children during the time surrounding meals (e.g., Goodenough, 1931). The further limitation of distractions increased the likelihood that the observational sessions would generate enough conflict behavior for hypothesis testing. Next, they tested observer influences on the data to identify whether any adjustments to their protocol were needed (e.g., they examined if “warm-up” sessions were necessary for families to adjust to having observers in their homes; results indicated that such accommodations were unnecessary; Patterson, 1982; see also Thornberry & Brestan-Knight, 2011). They examined the frequency of key behaviors and determined how much observation over how many sessions were needed to get a stable index of the behaviors of interest. They found that 60–100 minutes of data sampled in 5-minute blocks over the course of several sessions provided minimally stable estimates of boys' coercive behaviors. Finally, by using observations in a multitrait, multimethod assessment strategy (e.g., parent reports, global observer impressions, and school or arrest data), OSLC provided evidence for the validity of their observational approach and their coding system (e.g., Patterson, Reid, & Dishion, 1992).

Analogue Observation

Although naturalistic observation might be appealing because the required inferences about generalizability are minimized, analogue situations are often preferable because of their efficiency. Social psychologists employ analogue situations to (a) create environments where otherwise difficult or impossible to observe behaviors occur (e.g., how positions of power can evoke degradation, Haney, Banks, & Zimbardo, 1973); (b) enable observation of dynamic qualities of social interaction (e.g., escalation and de-escalation of negativity in mother-child dyads; Snyder, Edwards, McGraw, Kilgore, & Holton, 1994); and/or (c) isolate determinants of behavior.

An example of an observational paradigm of this type is the couples' problem-solving discussion (Heyman, 2001). Investigators typically ask couples to discuss one or two potential conflict areas for 10 to 15 minutes each. Within these general parameters there is wide variability in exactly how conversations

are structured. A prototypical protocol is presented in [Table 14.2](#). Other approaches include providing couples with standardized topics to role play (e.g., planning a vacation), which may not relate to their own conflicts (e.g., Aron, Norman, Aron, McKenna, & Heyman, 2000), or having them reenact prior conflicts (e.g., Margolin, Burman, & John, 1989). Other researchers have set up situations to observe couples providing social support (e.g., Pasch & Bradbury, 1998), sharing exciting activities (e.g., Aron et al., 2000), or discussing situations of high import (e.g., Schmalings, Wamboldt, Telford, Newman, Hops, & Eddy, 1996).

Table 14.2. *Dyadic Conflict Discussion Protocol*

-
1. Setup (prior to first interaction):
 - a. Check random number list to determine if the topic from Participant 1 (e.g., woman) or Participant 2 (e.g., man) topic is first.
 - b. Look at each participant's top areas of conflict (e.g., from Areas of Change Questionnaire). Pick top area of desired change for participant who will initiate first conversation. In case of a tie within a person, use random number sheet to determine order. If both participants pick the same topic, use it for whoever is randomly chosen to go first. Then choose the next-highest topic for the second participant's discussion.
 2. Instructions for conversations are given separately to participants (i.e., they are in different rooms):
 - a. To the participant who will initiate the discussion, begin with "You wrote that you'd like to see [other participant's name] change [conflict topic]..."
 - b. To the other partner, begin with "Your partner wrote that s/he'd like to see you change [conflict topic]..."
 - c. "We'd like you to have a conversation with [name] about that topic for 10 minutes and try to get somewhere with it. We'd just like to see you discuss this like you typically talk about problems when you

are [at home/in your dorm room/etc.]. [pause for questions] OK, we're just about ready. The last thing is to make sure that you know how you will start. Think to yourself about what you would do if you were to bring up [conflict topic] [at home/in your dorm room/etc.]. Do you know how you would start?"
[Check to make sure that she have some way to start]

3. Prior to second interaction:

- a. To the participant who will initiate the discussion, begin with "You wrote that you would like to see to see [other participant's name] change [conflict topic] ..."
- b. To the other partner, begin with "Your partner wrote that he/she would like to see you change [conflict topic]..."
- c. To both: "We'd like you to have a conversation with [name] about that topic for 10 minutes and try to get somewhere with it. Like last time, we'd just like to see you discuss this like you do at home/in your dorm room/etc.)."

Perhaps surprisingly, asking couples to engage in communication about conflictual topics while researchers watch tends to elicit behavior with reasonable external validity. First, observed conflict behaviors in home and laboratory settings tend to be similar, although lab conflicts are a bit less negative (e.g., Gottman, 1979; Gottman & Krokoff, 1989). Second, couples judge in-lab behavior as typical of what they do at home (Foster et al., 1997). Third, partners' reactivity and self-consciousness while being observed are relatively low (Christensen & Hazzard, 1983; Jacob, Tennenbaum, Seilhamer, Bargiel, & Sharon, 1994). Thus, even if in-lab "conflicts on command" are not quite as negative as they are at home, they still reveal detectable differences in affect, behavior, physiology, and interactional patterns and processes (e.g., Gottman, 1979, 1994, 1999).

Experimental Manipulation

Social psychologists study behavior within controlled laboratory settings to (a) observe behaviors that are not likely to be observed in unstructured settings and/or (b) to experimentally manipulate the causes of those behaviors. By controlling all aspects of a laboratory environment except that which is being manipulated, psychologists are able to isolate particular behaviors of interest and make conclusions about the cause of behaviors – an integral step to theory development (see Smith, [Chapter 3](#) in this volume). In addition, often in naturalistic settings there are multiple causes of behaviors that are interdependent, making it difficult to isolate which of several factors actually cause the behavior. With experimental manipulation, researchers can tease apart these causes by systematically manipulating them.

There are several issues to consider when designing an experiment in which the goal is to change behavior. Whether the manipulation is minimal or large and the degree to which behaviors are “difficult or easy to influence” are important considerations (Prentice & Miller, [1992](#), p. 162), and they are certainly relevant for studies that intend to influence the display of dynamic, interpersonal behaviors. Minimal manipulations that have large effects on behaviors can be particularly convincing in demonstrating the strength and size of an effect. The mere exposure effect and the minimal group paradigm are classic examples of minimal manipulations that produce large effects on behavior. As a more recent example, Goff, Steele, and Davies (2008) demonstrated that white participants who were led to believe that they would discuss racial profiling with an African-American participant placed their chairs farther apart from their partners’ chairs than did whites who were led to believe that they would discuss a race-neutral topic. Goff and colleagues’ manipulation is minimal because the mere belief that participants would have a race-based discussion was sufficient to alter behavior.

It is also important to consider the behaviors that are manipulated and measured. It is provocative to demonstrate that an experimental manipulation affects behaviors that are “difficult to influence” (Prentice & Miller, [1992](#), p. 162), largely because easy-to-influence behaviors are mundane (e.g., ask participants to sit when they arrive and they sit) and of little interest. Rapport-building within cross-race interactions and conformity to groups (e.g., Asch, [1951](#)) are examples of difficult-to-influence behaviors. Manipulations that are both minimal in nature *and* exert effects on such behaviors are often deemed particularly impressive by social psychologists, and are therefore more likely to make a scientific impact.

Studies that examine dynamic interpersonal behaviors, such as mimicry (Van

Baaren, Janssen, Chartran, & Dijksterhuis, 2009), self-disclosure (Altman & Taylor, 1973), or rapport-building (Tickle-Degnen & Rosenthal, 1990), require at least two individuals. As such, one of the most important methodological choices that social psychologists make in terms of experimental manipulations within social interaction studies is whether or not to use a confederate, rather than another participant, as the social interaction partner of interest. If the theoretical question of interest is interpersonal in nature – that is, it involves manipulating and measuring the behaviors of both partners within the interaction or examining the interdependence between partners' behaviors – one should strive to design a study in which real participants, not confederates, are used. However, there are a variety of situations where confederates may be appealing.

First, confederates offer a great deal of experimental control within an interpersonal interaction and are ideal in examining theoretical questions that are intrapersonal in nature. For example, Lakin, Chartrand, and Arkin (2008) examined how being socially excluded prior to a dyadic interaction influenced mimicry of the interaction partners' nonverbal behaviors. After receiving a social exclusion manipulation, participants interacted with a confederate who was trained to engage in a specific set of nonverbal behaviors, namely foot wiggling. The authors were interested in the degree to which participants who were socially excluded also wiggled their feet. In this example, the empirical question was *intrapersonal* – it involved examining how an individual-level predictor (social exclusion) influenced the behaviors of only one person in the interaction, not both. Second, confederates are a valid choice when the interaction partners' behaviors are the experimental manipulation. For example, Blascovich, Mendes, Hunter, Lickel, & Kowai-Bell (2001) had participants interact with a stigmatized other who had a large birthmark on her face – which was painted on using make-up – or with the one who had no birthmark. The presence of the birthmark was the experimental manipulation. Third, confederates allow for a clean standardization of the dependent behavior of interest. In Lakin, Chartrand, & Arkin (2008), for example, mimicry was clearly (and simply) defined as foot wiggling. Fourth, because confederates offer a level of experimental control that participant interaction partners do not, they allow researchers to isolate the causes of behavior to gain a better understanding of social processes.

There are a few important steps that must be taken when using confederates as social interaction partners. First, it is important to make sure that confederates are not a hidden source of variance. For pragmatic purposes, researchers often use two or more confederates in a study. These confederates might not always

behave consistently with each other, so one must make sure that the effect of the experimental manipulation does not depend on with which confederate participants interact. One potential method for addressing this issue is to treat the confederate (e.g., Amy the confederate versus Stacy the confederate) as a predictor of the dependent behavior of interest and as a moderator of the effect of condition on that behavior (making sure that the confederate is crossed with condition). Confederates can also be considered a source of variance in the analysis. Kenny, Mohr, and Levesque (2001) discuss methods for examining reliability of observers' judgments of participants' behaviors, many of which are applicable to studies that use confederates. For example, they discuss the importance of treating the observer as a source of variance – a method that can be easily adapted to treating the confederate as a source of variance.

Second, whenever possible, confederates should be blind to condition so that their behaviors are not inadvertently influenced. For example, Mendes, Major, McCoy, & Blascovich (2008) went to great lengths to ensure that the confederate did not know whether she had a birthmark painted on her face. Third, confederates may be trained to behave in a certain, consistent way across participants, but they might engage in automatic behaviors that are outside of their awareness, especially during social interactions, and these behaviors could influence the interaction. To make sure that confederates behave consistently across participants and across conditions, researchers should record the behaviors of confederates within each interaction, if possible – for example, by videotaping them and then coding their behaviors.

Sometimes confederates are used because they represent groups that are difficult to recruit to participate in research, either because they are not part of a convenience sample or because they are a small percentage of the sample population. In these cases, confederates serve a pragmatic purpose, even when the question of interest is interpersonal. For example, many cross-race interaction studies conducted in the United States have recruited White participants who then interact with African-American confederates. Although such a strategy allows the examination of cross-race encounters within the lab, it limits the understanding of cross-race interactions from the African-American perspective (Shelton & Richeson, 2006). As such, theories about the nature of cross-race interactions have become “one-sided” in that there is much cumulative knowledge about the attitudes and behaviors of whites but much less knowledge about the attitudes and behaviors of African Americans. This is just one example of how the use of confederates can have direct, and potentially profound, theoretical implications.

Behavioral Observation Coding Systems

Behavioral observation coding systems tend to be one of two types. *Topographical* coding systems measure the occurrence of behaviors. *Dimensional* coding systems measure the intensity of behaviors along a dimension (e.g., warmth, engagement). The choice of a coding system depends on the specific purposes of a study. Because of the costs involved in launching a behavioral observation enterprise, a system should be only as complicated as is necessary to fulfill the purposes of the research study. Ideally, one can find an existing system that meets the researcher's needs. As Bakeman and Gottman (1997, p. 15) noted, however, this choice should not be taken lightly: "We sometimes hear people ask: 'Do you have a coding scheme I can borrow?' This seems to us a little like wearing someone else's underwear. [Using] a coding scheme is very much a theoretical act, one that should begin in the privacy of one's own study, and the coding scheme itself represents an hypothesis, even if it is rarely treated as such."

Consequently, the researcher should begin with a set of hypotheses and design the coding system around these hypotheses. It is unfortunate when researchers realize *after* the coding has been completed that they failed to code a critical behavior. Given the need for specificity and completeness, a system should not be chosen without a researcher spending a significant amount of time observing and studying the phenomena of interest in a variety of ways. For example, a project might begin with a researcher watching and making observations with written or verbal notes over a period of several months. During the same period of time, a literature review can be conducted to find similar projects and to learn about what coding systems were used and how various practical issues were handled. Considerations of interest during this period of the research project include not only what to code but also when, in what settings, and by whom. As ideas narrow, pilot work will be required with practice participants and design modifications and changes will likely follow.

This background work might lead a researcher to discover that a "just right" coding system is simply not available. In this case, the researcher is in good company. Many coding systems are derivatives of past systems that were deemed in need of revision for various reasons. For example, the Family Interaction Coding System (Patterson, 1982) was developed during the 1960s for coding naturalistic family interactions in the home setting. This system was soon revised into the Marital Interaction Coding System (Weiss & Summers, 1983), which was developed for coding couples' problem-solving interactions in

a laboratory setting. Like Latin, these two “dead coding languages” are the source for dozens of offshoots (Kerig & Baucom, 2004; Kerig & Lindahl, 2000). Thus, a first step in developing a new system is to try to find a past system that is closest to what is needed and revise from there. The advantage of using an existing coding system (or a close derivative of one) is that much psychometric work on reliability, interrater agreement, and validity has already been conducted. The disadvantage, as just noted, is that existing coding systems might not be a good match for one's hypotheses.

Coding Units

The most fundamental property of a coding system is the sampling strategy for behavior, otherwise known as the coding unit. Coding units divide an observation into discrete segments, and each segment has the opportunity to be assigned a code, should one apply. The major sampling strategies employed in behavioral observation (see Table 14.3) are event, duration, interval, and time. Each strategy yields a different type of coding unit. Advantages and disadvantages of each of these strategies are discussed in Bakeman and Gottman (1997) and Haynes and O'Brien (2000). With each strategy, data richness and quality (e.g., retaining the sequential unfolding of events, reliability, validity) must be weighed against practical issues (e.g., expense, time, availability or practicality of recording devices, difficulty obtaining reliability). As noted by Margolin *et al.* (1998), even when a coding unit is presumably clear, technical issues, such as the quality of the audio track on a video recording or the speed of turn taking in an interaction, can make detecting some units difficult. This is one reason why researchers interested in verbal communication often create written transcriptions that are used together with audio and video feeds when coding.

Table 14.3. Coding Units

Sampling Unit	What Is Recorded	Example
Event	The occurrence of each behavior of interest.	Noting each time a smile occurs over a 10-minute recording period.
Duration	The length of each behavior	Noting the total length

Duration	The length of each behavior of interest (behavior onset and offset times).	Noting the total length of time smiling occurs over a 10-minute recording period.
Interval	The occurrence of each behavior of interest in each consecutive time block/interval.	The presence/absence of smiling is noted for each 5-second interval during a 10-minute recording period.
Time	Intermittent observations, typically using duration or interval sampling, and the occurrence (and sometimes the frequency) of behaviors of interest.	Using event, duration, or interval sampling of smiling but only every other minute during a 10-minute recording period.

Molar versus Molecular Approach

Another key property of a coding system is how often codes are recorded. In molar, or “global,” coding systems (e.g., Rapid Couples Interaction Scoring System; Krokoff, Gottman, & Haas, 1989) summary ratings are made for each code over a large number of potential coding units (e.g., every 3 minutes in a 15-minute observation, or once at the end of the observation). Codes tend to be few, representing behavioral classes (e.g., negativity, attentiveness, escalation, reciprocation). Thus, numerous examples of the codes of interest may occur within multiple potential coding units, but only one summary score is given, usually indicating the frequency with which a code appeared throughout the observation period.

In contrast, molecular, or “microbehavioral,” systems code behavior as it unfolds over time and tend to have many fine-grained codes (e.g., eye contact, criticize, whine, withdraw) that are given within each coding unit. The large number of codes in many microbehavioral systems may make them inefficient to use, even with highly trained coders. This is because (a) coders can almost never get or maintain adequate interrater agreement on such a large number of codes, and (b) the codes occur too infrequently in a limited observational period to make them all useful even if they were reliably coded. Thus, researchers often

resort to grouping codes, often condensing down a large system into positive, negative, and neutral classes for analytic purposes (see review in Heyman, 2001). Imagine spending the extreme time and expense required to train coders on 40 codes, only to end up only analyzing positive, negative, and neutral!

Microbehavioral systems tend to be topographical; global systems can be either topographical or dimensional, though dimensional coding, especially on a behavior-by-behavior basis, is less common. Given that many theoretical models of interest have implicit or explicit intensity X time predictions (e.g., Patterson's [1982] Coercive Family Process model posits that reinforcement of escalating negativity contributes to the development of antisocial behavior in boys), this is unfortunate.

Noting drawbacks of microbehavioral coding systems (e.g., time to code and train, need to combine micro codes into categories, difficulty achieving interrater agreement) but wanting to retain the advantages (e.g., specificity, sequential relations), researchers (e.g., Gottman, 1996; Heyman, 2004) began developing a new generation of coding systems containing codes that could be analyzed without resorting to massive agglomeration (e.g., categories such as “hostility” instead of separate codes for negative voice tone, hostile content, eye rolls, etc.). Some of this work was guided through statistical analyses rather than a priori decisions. For example, Heyman, Weiss, Eddy, and Vivian (1995) factor-analyzed observations of more than 1,000 couples that had been coded with a 40-code system to derive a system that could code at a categorical level, thus streamlining training, coding, and analysis. The resulting system (Heyman, 2004) was still microanalytic but was much more practical to use (and had better reliability and validity) than the coding system it replaced (see also Whaley, Pinto, & Sigman, 1999).

Global systems are simpler and faster and can sometimes represent the construct of interest better (e.g., an overarching construct such as overreactive parenting may be better coded with a global code, where context can better be taken into account, than with microbehavioral coding). Some constructs, such as progress made in a problem-solving task, can only be coded globally. However, agreement can sometimes be difficult to obtain because of the lack of anchoring of ratings to specific behaviors. Furthermore, global systems do not maintain sequential relations, making them less useful for analyzing patterns (unless the coders specifically coded for that pattern). In an effort to obtain the “best of all worlds,” microanalytic and global systems have been paired, usually by asking coders to make global impressions ratings after coding microanalytically (e.g.,

Patterson, Reid, & Dishion, 1992) .

Multiple Dimensions

Another property of a coding system is how many different dimensions of an interaction are coded, and how many different codes are included within each dimension. For example, some coding systems record information about the general context within which a behavior is occurring (e.g., in a system focused on child behavior at school, this would be the location of an interaction, such as on the playground, in the lunchroom, or in the classroom), as well as the specific behaviors of interest. Other systems might also include a code describing the quality of the behavior, such as whether it was delivered with negative, positive, or neutral affect. The choice of how many dimensions to code depends on the specific hypothesis of interest, but issues can get confused in no small part because of the high cost of conducting observational work. Once data have been collected and a team has been assembled, it may seem appealing to collect as much information as possible while coding so that a variety of tasks can be accomplished, from hypothesis testing to hypothesis development. The most obvious risk in such an approach is increased difficulty in reaching an acceptable level of interrater agreement, but it may overly burden coders and compromise even more important qualities, such as the reliability and/or validity of the observation. This can only be known if other types of data (from multiple sessions, from multiple informants, through multiple methods) are collected to aid in understanding the observational data that is collected.

Example

An example of a mature coding system is the Interpersonal Process Code (IPC; Rusby, Estes, & Dishion, 1991), a distant tributary of the aforementioned Family Interaction Coding System. In the IPC, a target individual is chosen as the focus of an observation, and everything that individual does, and has done to him/her, is coded. The coding unit is a codeable behavior, which can continue even when the behaviors of others are also taking place (e.g., a target child starts to hum and continues to hum, even though the child he is playing with is yelling at him). When no codeable behavior is occurring, a Stop code is entered. When an individual cannot be fully heard or seen, an Out of View code is given. The IPC is coded on a handheld or stationary computer in real time and has been used to code both live and videotaped sessions.

Three dimensions are coded simultaneously in the IPC: Activity, Content, and

Valence. Activity refers to the general context within which a social interaction is taking place and varies depending on the study. An example of Activity codes used in prior studies is Work, Play, Read, Eat, Attend, or Unspecified. Activity codes are given in priority so that if a code with theoretically higher priority occurs, it is given (e.g., Work trumps Play, Play trumps Read, etc.). Content refers to specific behaviors of interest. Thirteen Content codes constitute the IPC and include positive, neutral, and negative verbal, nonverbal, and physical codes. For example, the code Positive Interpersonal is assigned to “verbal expressions of approval of another's behavior, appearance or state” (p. 17). Valence refers to the emotion tone accompanying the delivery of content (i.e., Happy, Caring, Neutral, Distress, Aversive, and Sad). In addition, who displayed the behavior (the Initiator) and whom the behavior was directed toward (the Recipient), are also coded.

Training Observers

The careful training of observers (i.e., “coders”) is essential to behavioral observation. People who may have very different perceptions of behavior must, through the training process, come to be interchangeable with one another. Moreover, they must maintain consistency over time. By analogy, two different watches should show the same time. Over time, the watch's estimates should remain unchanged. Of course, the social judgments made by human observers are known to be fallible and certainly less precise than a watch. Thus, interrater agreement should be meticulously attended to. Failure to do so can result in increased error variance, which constrains reliability and hampers the capacity to find associations of the coded behavior with other factors (even if they truly exist).

Coder training will be covered in a somewhat cursory fashion here, having been described in more detail elsewhere (e.g., Bakeman & Gottman, 1997). We will use the example of making ratings from video recordings, although the principles are broadly applicable to coding live or from audio only. The first phase of training involves familiarizing the coders with the constructs being measured and the observational context (being careful not to reveal study hypotheses) and the reasons for the heavy focus on obtaining interrater agreement. A manual should (a) describe the coding procedures in detail, (b) clearly spell out distinctions among behaviors, and (c) provide illustrative examples. Because a video example is worth a thousand words, the trainer should have an ample supply of video clips illustrating the behaviors. We

recommend beginning with cardinal examples that are relatively easy to discern. Over the course of training, the examples should get progressively more challenging, illustrating finer distinctions. Meetings are typically held two to three times per week with the trainer and all coders present. Between meetings, coders practice on a carefully selected set of video recordings. The training videos should be selected to illustrate the full range of behaviors being coded, with progressively more difficult cases presented over time. Meetings are used to review the process of coding (e.g., the reasoning behind coding decisions) and clarify decision rules and sources of disagreement. Interrater agreement is calculated for each video assigned and reviewed in the meetings. This phase lasts for as long as necessary to achieve sufficient agreement. For categorical data, we suggest training coders until they consistently agree with the ratings of a master coder about 70% to 90% of the time, depending on the complexity of the coding system. For dimensional ratings, we suggest training coders until the majority of scores are in point-by-point agreement with a master coder, with disagreements very rarely greater than one point. These strategies will usually well exceed standard benchmarks for acceptable interrater agreement (i.e., *Kappa* or *AC1* above 0.6, *ICC* above 0.7; see “Interrater Agreement” section later in the chapter).

On reaching the aforementioned criteria, the coders are ready to begin producing “real” data, and there is a shift from training to maintenance. In the most typical case, the videos to be coded are divvied up among coders, with only partial overlap in which videos are coded by two or more different people. These overlapping cases are used to assess interrater agreement (to be reported in resulting manuscripts); thus, it is crucial that the coders are not informed about which videos will be used for assessing agreement. Additionally, it is important that these “reliability videos” are selected at random, typically during each week of coding. After all coders have completed the reliability videos each week, they are then reviewed in meetings and the reliability statistics are presented to the coders. The purpose of these meetings is to maintain coders’ performance and prevent shifts in the use of rating criteria over time. Typically, the reliability sample consists of a randomly selected 25% to 33% of all videos coded.

Interrater Agreement

Clearly, interrater agreement is an important consideration when coding. One must be able to establish that the codes recorded from an observation are not just one person's idiosyncratic view of the world but reflect a standard, albeit

imperfect, set of definitions that can be applied with nearly identical results by other people and in the same manner across time. Interrater agreement statistics are quantitative aids for this task. They clearly are useful and vital in training coders, where interrater agreement statistics can be used to monitor progress toward a quantitative agreement criterion. Interrater agreement statistics can also be usefully employed to monitor and correct drift in the use of rating criteria once coders move beyond training. Finally, reporting of interrater agreement in published works is important to help readers evaluate a study's methods and findings.

Although “interrater agreement” and “reliability” are sometimes used interchangeably, this is sloppy usage. Mitchell (1979) provides a cogent discussion of the critical distinctions, grounded in classical measurement theory. To summarize, reliability reflects the degree to which variability in obtained scores (e.g., the ratings assigned by a coder) reflects variability in the underlying trait being measured. Interrater disagreements reflect only one of several threats to reliability; others include random fluctuations in subjects' behavior, in the setting, and in the protocol. Thus the degree of interrater agreement is not equivalent to the degree of reliability in the measure of behavior, as it is only one piece of the pie. However, because interrater agreement is controllable by the investigator, it has received the most attention. As disagreements increase, measurement error increases and reliability and validity decrease .

Which Interrater Agreement Statistic to Use

The scale of measurement and the variables formed from observational data largely determine the interrater agreement statistic used. One set of statistics is most appropriate for categorical or nominal judgments (e.g., deciding which of several emotions a person is expressing), which tend to be the basis for molecular, or microbehavioral, coding systems. Another set is more appropriate for ordinal, interval, or ratio scale judgments (e.g., deciding how intense a person's expression of a given emotion is), which are sometimes the basis for global coding systems. The distinction is not absolute, however, depending on the intended usage of the observations. For example, categorical ratings are often summarized across an entire observation period into frequency and duration variables, in which the latter set of tools can be applied; however, Bakeman and Quera (2011) maintain that it is still important to establish behavior-by-behavior (i.e., local) agreement in these contexts. In contrast, if behavioral sequences are to be analyzed, then establishing behavior-by-behavior agreement would be

required, not optional. In the following sections we describe the most common interrater agreement statistics, as well as some useful alternatives.

Categorical Observations

Interrater agreement statistics for categorical observations each begin with the raw proportion of agreement between raters. Yet, even people who make ratings purely at random agree with one another some of the time by pure chance. Accordingly, interrater agreement statistics adjust for this possibility, with the differences among these adjustments responsible for the differences between the statistics.

The frequencies of agreement and disagreement are helpfully represented in what is known as a confusion matrix (see [Table 14.4](#)). A simple confusion matrix is presented for the situation in which two coders' agreements and disagreements in the presence versus absence of a given behavior are represented. Agreements are found in bold along the diagonal, with "a" representing the number of agreements on the presence of a behavior and "d" representing the number of agreements on the absence of a behavior. Disagreements are found in the off-diagonal cells, "c" and "b." The row ("e" and "f") and column ("g" and "h") totals are referred to as marginal frequencies or simply marginals; they represent the frequencies for each coder's ratings of behavior presence and absence. Each of the interrater agreement statistics are calculated from the tallies in the confusion matrix.

Table 14.4. Confusion Matrix for Presence vs. Absence of a Behavior Rated by Two Coders

Coder 1	Coder 2		Row (Coder 1) Totals
	Behavior Present	Behavior Absent	
Behavior Present	a	b	$a + b = e$
Behavior Absent	c	d	$d + e = f$
Column (Coder 2) Totals	$a + c = g$	$b + d = h$	$a + b + c + d = i$

Cohen's Kappa. Cohen's (1960) kappa is probably the most widely used interrater agreement statistic and is given by the following formula, with reference to the [Table 14.4](#) example of a single code's presence or absence rated by two observers:

$$\kappa = (P_o - P_{e|\kappa}) / (1 - P_{e|\kappa}),$$

where P_o is the observed agreement, found along the diagonal of [Table 14.4](#) and given by

$$P_o = (a + d) / i,$$

and $P_{e|\kappa}$ is the kappa model expected or chance agreement, calculated by considering the row and column marginals and given by

$$P_{e|\kappa} = [(e^*g) / i + (f^*h) / i] / i.$$

Kappa generalizes to accommodate multiple codes; however, code-by-code interrater agreement is essential to establish and report, as disagreements on a code can go unnoticed if embedded in a larger matrix with other codes for which there is better agreement. Moreover, kappa tends to be larger with a greater number of codes (Bakeman, Quera, McArthur, & Robinson, 1997), potentially yielding overly optimistic estimates of interrater agreement.

Kappa is straightforward to calculate (by hand or by using spreadsheets such as Excel), but it can also be calculated in standard statistics programs (e.g., SPSS) and with Robinson and Bakeman's (1998) ComKappa program; the 2010 update is available for download from Bakeman's website: www2.gsu.edu/~psyrab/BakemanPrograms.htm. Kappa can also be calculated with the “irr” package in the free statistics program, R (cran.r-project.org/web/packages/irr/irr.pdf; R Development Core Team, 2005).

By far the greatest limitation of kappa is how it is affected by distributional asymmetries (i.e., high or low rates of a given behavior). These distributional asymmetries are referred to as skewed marginals because the row or column totals or marginals (“e” vs. “f” or “g” vs. “h” in [Table 14.4](#)) are lopsided, as they tend to be in psychological research. This is in large part because – as noted earlier by Festinger and Mehl – many of the most interesting psychological phenomena are relatively infrequent compared with the mundane. The effect of skewed marginals can be seen in [Figure 14.1](#). Panel A models the performance of Cohen's kappa, as well as the other statistics in this section given 90% interrater agreement, with evenly apportioned disagreements in the presence versus absence of the behavior being rated. When a behavior occurs at a rate of 50%, and the marginals are thus perfectly balanced, kappa is .80 (i.e., a 10% downward adjustment for random agreement). The greater the deviation from

balanced marginals, the greater the adjustment. When the behavior reaches an 80/20 split (i.e., present or absent 80% of the time), kappa is .69 (i.e., a 21% chance agreement adjustment). The adjustment is even greater when behaviors are very rare or very frequent, with kappa falling to .44 at a 90/10 split. Panel B shows that kappa's sensitivity to skewed marginals is even greater at 80% interrater agreement.

Typical rules of thumb for interpreting kappa and similar statistics are that kappas from .40 to .59 are fair, .60 to .74 are good, and .75 and above are excellent (Cicchetti, 1994). Yet the skewed marginal problem severely challenges these guidelines (Bakeman et al., 1997), as it is nearly impossible to achieve substantial kappas with highly skewed data. Gwet's (2008) Monte Carlo analyses demonstrate that kappa's chance agreement is incorrect for very common or uncommon behaviors, thus decreasing the utility of kappa in a very common research scenario.

Weighted Kappa. Weighted kappa (Cohen, 1968) is an alternative to kappa that allows the researcher to penalize more heavily for some disagreements than others. In contrast, unweighted kappa regards all disagreements as equally serious. The weighted kappa is rarely used with nominal data in social psychology, perhaps because of the difficulties in establishing and convincing others of the validity of the weights (Bakeman & Quera, 2011). However, it involves creating a weights matrix that specifies the severity of disagreements. For example, one might decide that disagreements in rating different forms of negative emotion are less serious than disagreements in rating negative versus positive emotions. This might lead one to weight anger vs. contempt disagreements as 1 and happiness vs. anger or contempt difference as 2 in the “weights matrix” (i.e., a grid containing the weights for all possible combinations of ratings of the two coders being compared). Agreements are assigned 0 in the weights matrix. The weights are simultaneously taken into account, alongside the observed and expected or chance agreements in calculating the weighted kappa. Weighted kappa is given as:

$$\kappa_w = 1 - \frac{\sum_{i=1}^k \sum_{j=1}^k w_{ij} x_{ij}}{\sum_{i=1}^k \sum_{j=1}^k w_{ij} e_{ij}},$$

where k is the number of codes and w_{ij} , x_{ij} , and e_{ij} correspond to elements (i -th row and j -th column) in the weight, observed, and expected matrices, respectively. Borrowing Bakeman and Quera's (2011) notation, $e_{ij} = p_{+j}x_{i+}$ with

x_{i+} the sum of the i -th row, p_{+j} the probability for the j -th column, and $p_{+j} = x_{+j}/N$.

Fortunately, weighted kappa can easily be computed using spreadsheets such as Excel, with ComKappa or with the “irr” package in R (see earlier discussion) .

Van Eerdewegh's V. Spitznagel and Helzer (1985) offer a statistic called V as an alternative to kappa, which is less sensitive to the skewed marginal problem. We refer to this statistic as Van Eerdewegh's V (after the statistic's author), to distinguish it from Cramér's V (Cramér, 1946). V is given by:

$$V = \left(\sqrt{a*d} - \sqrt{b*c} \right) / \sqrt{(a+c)*(b+d)}$$

As seen in Figure 14.1, V is identical to kappa with balanced marginals, with greater differences at greater splits, in which V is always larger than kappa. With 90% agreement and a 90/10 split in the marginals, however, V (.52) is only slightly larger than kappa (.44), as seen in Panel A. Thus, V , like kappa, is sensitive to skewed marginals. As with kappa, this sensitivity is even greater at lower levels of interrater agreement (see Panel B). In sum, V is only slightly superior to kappa as a metric of interrater agreement for very common or uncommon behaviors. Its performance has not yet been evaluated in Monte Carlo simulations, to our knowledge.

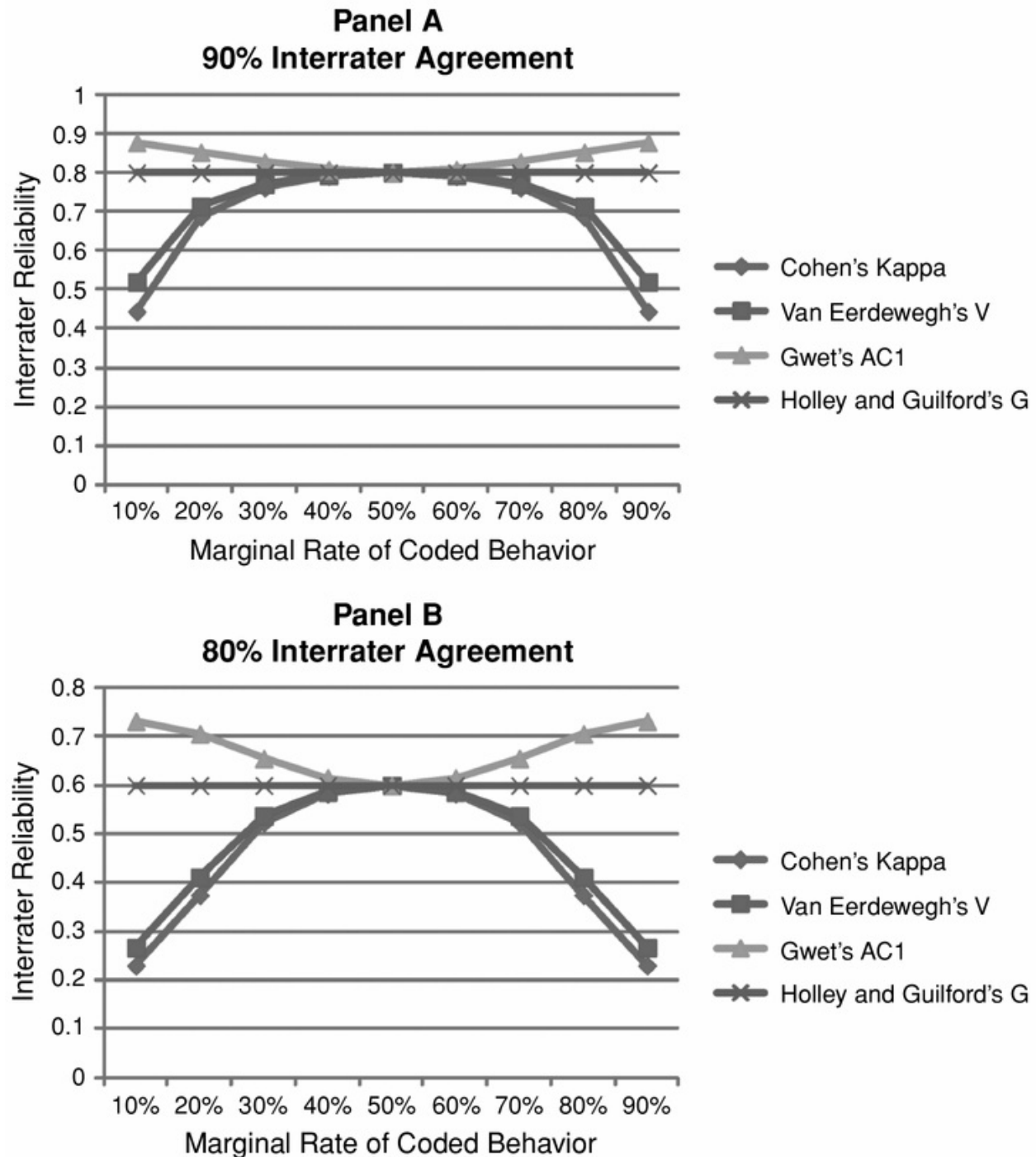


Figure 14.1. The performance of five interrater agreement statistics across different marginal rates of behavior, and for 90% (Panel A) and 80% (Panel B) raw interrater agreement.

Holley and Guilford's G. Holley and Guilford's (1964) *G* is like kappa, except in the manner in which chance agreement is calculated. In contrast to kappa, in

which chance agreement varies with the marginal rates of behaviors, G assumes a fixed rate of chance agreement. A generalized form of the equation is given by Gwet (2008):

$$G = (P_o - P_{e|G}) / (1 - P_{e|G}),$$

with $P_{e|k}$ the G model expected or chance agreement, given by:

$$P_{e|G} = 1/q,$$

with q equal to the number of response categories. In present/absent comparisons, chance agreement is always .50. Thus, G has zero sensitivity to skewed marginals, resulting, for example, in a value of .80 with 90% agreement and .60 with 80% agreement (Figure 14.1). According to Gwet's (2008) Monte Carlo simulations, G is less biased than kappa.

AC1. Gwet (2002) recently developed the $AC1$ statistic as an alternative to kappa, which is less sensitive to the skewed marginal problem of kappa, and is given by:

$$AC1 = (P_o - P_{e|AC1}) / (1 - P_{e|AC1}),$$

with P_o identical to kappa, and with $P_{e|AC1}$ the $AC1$ model expected or chance agreement:

$$P_{e|AC1} = 2 * P_+ (1 - P_+),$$

where $P_+ = [(e + g)/2]/i$, and with reference to Table 14.4. As seen in Figure 14.1, $AC1$ is identical to kappa, V , and G with balanced marginals. Greater differences between $AC1$ and the other metrics emerge at greater splits. Notably, $AC1$ assumes somewhat *lower* chance agreement with skewed marginals, in contrast to kappa's opposite assumption. To illustrate, at a 90/10 split and 90% agreement (Figure 14.1, Panel A), $AC1$ is .88 (compared with a kappa of .44), whereas it is .80 with a 50/50 split. Thus, $AC1$ does not over-penalize one for skewed marginals. Gwet's (2008) Monte Carlo simulation data suggest that $AC1$ produces significantly less biased estimates of interrater agreement than does kappa, and it slightly outperforms G , as well.

Summary and Recommendations. The categorical interrater agreement statistics we have presented produce comparable results when behaviors are neither very frequent nor infrequent – any metric will do in such situations.

However, behavioral observations are frequently skewed, and metrics other than Cohen's kappa have been shown to be superior. *G* and *AC1* stand out from kappa and Van Eerdewegh's *V* in this regard, and *AC1* produced less biased estimates of interrater agreement than *G* in Gwet's (2008) Monte Carlo simulations. Thus, we tentatively recommend the *AC1* as the preferred metric. Caution is warranted, however, as it has not been used widely and has only been evaluated in a single Monte Carlo analysis. Additional study is warranted. Moreover, there are no established rules of thumb for what constitutes, for example, poor and good *AC1*s. However, we believe that it would be reasonable to apply the long-standing criteria for judging kappa (i.e., .60 to .74 is good and .75 and above is excellent; Cicchetti, 1994), with no allowances made for distributional characteristics.

Ordinal, Interval, and Ratio Observations

Intraclass correlation (ICC). The use of the *ICC* for interrater agreement in psychology was popularized by Shrout and Fleiss (1979) and is widely applied to ordinal, interval, and ratio scaled observations – although technically it assumes interval or ratio data. In simple terms, the *ICC* parses variation in observers' ratings into (a) variance owing to differences among the subjects being observed and (b) variance owing to the observers. Interrater agreement, hence the *ICC*, increases to the extent that between-subject variance is greater than between-observer variance, couched in familiar ANOVA terminology. The *ICC* takes into account disagreements in the rank ordering of subjects, as well as the means and the variance. To illustrate, if Coder X rates subjects A, B, and C as exhibiting a mean anger intensity of 1, 2, and 3, and Coder Y rates them as 3, 4, and 5, respectively, the *ICC* will punish the disagreement in means (2 vs. 4, respectively), yielding an *ICC* of zero, despite perfect agreement in the rank ordering of the subjects. This example also clearly shows the inadequacy of the Pearson and Spearman correlations for judging interrater agreement, as they are 1.00 despite no absolute agreement in the behavior being rated.

There are several different versions of the *ICC*, raising questions about which to use. Each is estimated in the context of ANOVA, in which variance is segmented into different parcels, such as between subjects (i.e., variation among the people being rated) and within subjects (i.e., variation among the raters). Each easily generalizes to more than two raters.

The most useful *ICCs* for assessing interrater agreement treat coders as a “random effect,” meaning that the set of coders used in a given study has been

randomly selected from a larger population of coders (Shrout and Fleiss (1979)). It is rarely the case that the particular set of coders who assist us in our research are the only coders of interest and to whose ratings in future studies we would like to generalize; such would be a case for fixed effects analysis.

Whitehurst (1984) suggests the use of a one-way random effects ANOVA, given as:

$$ICC = [MS_B - MS_W] / [MS_B + (k - 1) * MS_W],$$

where MS_B is mean square between subjects, MS_W is mean square within subjects, and k is the number of judges.

Other writers (e.g., Bakeman & Quera, 2011) suggest the use of one type of two-way random effects ANOVA; see McGraw and Wong (1996) for others. The primary difference from the one-way approach is that MS_W is subdivided into its components, MS_E (mean square error) and MS_O (mean square observer). Furthermore, there are two different versions of the two-way random effects ICC . The first is called the relative consistency ICC and is given by:

$$ICC_{rel} = [MS_B - MS_E] / [MS_B + (k - 1) * MS_E].$$

The second version is called the absolute agreement ICC and is given by:

$$ICC_{abs} = [MS_B - MS_E] / [(MS_B + (k - 1) * MS_E) + k/n * (MS_O - MS_E)],$$

where n is the number of subjects. With the two-way random effects approach, we recommend the absolute agreement ICC , the more stringent of the two, in that it reflects more than just whether coders provide similar rank ordering of the behaviors being rated (i.e., relative agreement), but also the degree to which the coders are interchangeable – the highest proof of agreement. The one-way approach is similarly stringent.

Cicchetti's (1994) review suggests the same rules of thumb for interpreting $ICCs$ as for categorical metrics (e.g., kappa). However, we recommend a higher criterion: acceptable $ICCs$ should exceed .7, and .8 and above is very good. In our experience, $ICCs$ below .7 often result from multipoint discrepancies between coders (which suggests the need for more training) or from skewed distributions (which suggests the need to use a different statistic).

Unfortunately, the *ICC* suffers a problem similar to that of Cohen's kappa. The *ICC* is compromised by skewed distributions. As pointed out by Whitehurst (1984), the *ICC* assumes a normal underlying distribution of the trait being measured, with deviations from normality resulting from rater error. However, many variables of interest in social psychology can be expected to be skewed. To return to the example of anger intensity ratings, unless an experimental manipulation is unusually powerful, most subjects can be expected to exhibit lower levels of anger, with fewer and fewer subjects showing higher levels of anger – they will likely be positively skewed. Accordingly, the *ICC* is not always the best choice.

The *ICC* can be calculated in common statistical packages (e.g., SPSS), as well as in the “irr” package in the free statistics program, R (cran.r-project.org/web/packages/irr/irr.pdf; R Development Core Team, 2005).

Finn's *r*. Finn's *r* (Whitehurst, 1984) is an alternative to the *ICC* that is less sensitive to skewed distributions. Also, whereas the *ICC* assumes interval or ratio scaled data, Finn's *r* assumes ordinal structure. This, too, is a positive feature in social psychological research in which observational ratings are often made on single scales that may have only 3, 5, or 7 points, and in which even intervals between scale points cannot be assumed to be even (failing to satisfy criteria for interval scale measurement) and/or do not have a meaningful 0 point (failing to satisfy criteria for ratio scale measurement). Unless such ratings are subsequently averaged (e.g., across multiple experimental periods, similar to items on a questionnaire), ordinal statistical models may be the best fit. Finn's *r* is given by:

$$r_f = (S_c^2 - S_0^2) / S_c^2,$$

where S_c^2 is the expected within-subjects variance when the ratings are assigned randomly and S_0^2 is the MS_W from a one-way random effects ANOVA with independent ratings of each subject as the within subjects variance. S_c^2 is given by:

$$s_c^2 = (k^2 - 1)/12,$$

where k is the number of ordinal scale categories.

As a rule of thumb, we suggest that Finn's *r* should be above .7 to be

considered acceptable. In our experience, however, Finn's r appears impracticably inflated with a greater number of scale categories, including when allowing half-points (i.e., 1, 1.5,...6.5, 7).

Finn's r can be calculated with the aforementioned formulas from quantities in the *ICC* output of common statistical programs or with the “irr” package in R (see *ICC* section earlier in the chapter).

Weighted Kappa. Weighted kappa, described in the prior section, can be an alternative to the *ICC* or Finn's r for ordinal data. As pointed out by Bakeman and Quera (2011), weighted kappa is more easily defensible for ordinal than for nominal data, because the weights assigned to disagreements are less arbitrary in the former case. Disagreements that are further apart on an ordinal rating scale should be penalized more heavily than those that are closer. For example, if aggression in a peer competition task is coded as absent, low, and high, disagreements between ratings of absent vs. high are more serious than absent vs. low or low vs. high disagreements. Linear weights are the most common in such applications. One-point disagreements are assigned a weight of 1, two-point disagreements a weight of 2, and so on. Quadratic weights, the square of linear weights, are also possible, penalizing far-apart differences even more heavily.

Summary and recommendations. The *ICC* is the only choice for truly continuous data. Finn's r is a solid alternative to the *ICC* when skewed distributions are a concern, with the proviso that it appears to be inflated when the number of scale points is high. Finn's r and weighted kappa are each viable choices for ordinal data.

Interrater Agreement for Sequences

Bakeman *et al.* (1997) point out that even reasonable levels of behavior-by-behavior interrater agreement does not guarantee that *event sequences* computed from these behaviors are in close agreement. Accordingly, the authors recommend a two-stage process in which interrater agreement is first computed at the level of the behavior, establishing local agreement. Next, the sequences of interest are computed, using one of the metrics of sequential association (e.g., Yule's Q for categorical data, and the lagged cross-correlation for continuous data). Each subject or dyad has such a value for each sequence of interest. These sequential association metrics are then compared for interrater agreement, using the *ICC*. This approach is very stringent and has not been used often (e.g., Martinez & Forgatch, 2001). Nonetheless, the simulation data of Bakeman *et al.* (1997) suggest that there can be significant degradation in the interrater

agreement of event sequences, compared with local interrater agreement. The two-stage process offers protection against this concern. Bakeman, Quera, and their colleagues offer software for determining event sequence interrater agreement (ELign; Quera, Bakeman, & Gnisci, 2007), which can be downloaded from www2.gsu.edu/~psyraab/BakemanPrograms.htm.

Reliability across Observations, Contexts, and Time

Generalizability theory (Gleser, Nanda, & Rajaratnam, 1972) is an extension of the statistical foundations undergirding the ICC. That is, the ICC is based on components of variance attributable to coders, targets of observation, and their interaction. Generalizability theory elegantly partitions variance attributable to multiple instances within a facet (multiple coders rating the same video) or multiple sources of variance (multiple coders and multiple observations). In the simplest use (multiple coders), Cronbach's alpha can be calculated for an event-coded system, with frequency counts for codes standing in for the “score” on that “test item” and coders standing in for test takers. As an example of using generalizability theory for multiple sources of variance, Wieder and Weiss (1980) partitioned variance attributable to (a) one, two, and four samples and (b) coders in (c) both audio and video conditions. The behavioral samples were collected three weeks apart. For both audio and video samples, most of the variance was accounted for by the “true variance” components (across people and across behavioral samples) and little by the error sources (coders, first vs. second samples, coder x people, coder x behavioral sample, or coder x people x sample).

How Much Time Is Necessary to Achieve Acceptable Reliability?

As noted earlier, investigators often use “reliability” and “interrater agreement” interchangeably. Yet, interrater agreement is but one component of stability of measurement (e.g., Hops, Davis, & Longoria, 1995; Kelly, 1977; Mitchell, 1979; Suen, 1988). It is also affected by the stability of the behaviors observed, which is highly dependent on the length of observation. By treating observation intervals as test items, Waters (1978) was able to use conventional psychometric statistics for reliability to determine how long one would have to observe to achieve a set level of reliability (i.e., stable results). In the psychometric theory of test reliability (Cronbach, 1951), test reliability can be assessed via the

coefficient of correlation between scores on comparable halves of the test (Ghiselli, Campbell, & Zedeck, 1981). When applying similar principles to observational data, Waters suggested that each 30-second sampling interval be considered a test item that is passed or failed (i.e., the target behavior occurs or does not occur). Interval-based coding systems would already be in a form for such statistics. Event-based coding systems can be converted into an interval-based system by using 30-second windows for the events coded. Time intervals can then be sorted into odd (1st, 3rd,...k) and even (2nd, 4th,...k-1) groups. The correlation between the odd and even group is the split-half internal consistency reliability for observed variable of interest. (Step-by-step instructions for calculation can be found in the appendix of Heyman et al., 2001.) Heyman *et al.* (2001) found that in couples' conflict observations, 10 to 15 minutes (the most typically used length, established through convention) was sufficient for stable estimates of most codes .

Validity

An important question related to behavioral observation is whether the variables generated are valid measures of the behaviors of interest. Unfortunately, there has been a tendency for researchers to assume that the variables generated from behavioral observation are somehow “more objective, less biased, or inherently superior” (Jacobson, 1985, p. 298) than other measures (such as self-report questionnaires), and this may have limited the examination of the validity of observational measures. Of course, whether or not a measure has more desirable properties than another measure is an empirical rather than philosophical question, and thus cannot be addressed unless data are collected.

The type of validity that is most often cited is face validity. Because behaviors are labeled, and often the labels are relatively straightforward, their validity seems self-evident (thus leading to comments like those of Jacobson). To increase confidence in these variables, however, more information is required. Probably the most important type of validity is construct validity (i.e., whether a tool truly measures what it is intended to measure). This is established via (a) convergent validity, or whether the observed variables (behavior or behavioral pattern) are associated with measures of the same construct that were collected by other means (e.g., “global” self-report, in-person interview, diaries, or reports of the past 24 hours); and (b) discriminant validity, or whether the observed variables are not associated with measures of different constructs. Predictive validity (whether a tool is related to future outcomes in a hypothesized manner)

is especially important in studies that observe behavior to predict outcomes longitudinally (e.g., if roommate conflict early in the year predicts GPA). Finally, discriminative validity (whether a tool can distinguish among groups that are hypothesized to differ) can be used both as a substantive test (e.g., do conflictual and nonconflictual roommate dyads differ on a measure of observed problem resolution?) and as a manipulation check (e.g., does the measure of obnoxious behavior differ in high and low confederates annoying conditions?)

Analyzing Behavioral Observation Data

When analyzing behavioral data, one must consider both *how* the behavior is measured and *how often* it is measured. In terms of *how*, for example, behavior can be measured continuously, such as by having observers record impressions of how anxious a participant appeared during an interaction using a 1 (not at all) to 7 (very much) scale. Behavior can be measured through a simple count – for example, by having observers log how often a participant blinked during an interaction. The relative nature of one behavior versus others can be measured – for example, the percentage of household labor completed by one partner in a romantic relationship is recorded relative with the other. Lastly, behavior can be measured dichotomously – for example, whether participants wore a condom during sex.

In addition to how behaviors are measured, it is also important to consider how often they are measured. In some cases, each behavior is only measured once for each participant so that data can be analyzed using traditional statistical methods such as regression or ANOVA for continuous outcomes, Poisson regression for count data, or logistic regression for dichotomous outcomes. However, in other cases, behavioral measures are collected several times using a repeated-measures design, or they are measured on several occasions over time. For example, during a 15-minute interaction, participants' behaviors may be recorded once per minute, for a total of 15 recordings per participant. In a daily diary study, participants might report on whether they had a fight that day with their romantic partner, for 15 consecutive days. In both of these examples, the data are multilevel because the behaviors of each participant are measured several times, and so time points (or repeated measures) are nested within participants. How participants behaved at one time or repeated measure is likely correlated with how they behaved at another time or repeated measure, and so the nonindependence in behaviors needs to be adjusted for (Kenny & Kashy, Chapter 22 in this volume).

There are several different analytical methods one might employ when analyzing multilevel behavioral data. One strategy that is optimal for many different types of outcomes – linear, count, and dichotomous – is General Estimating Equations (GEE; Liang & Zeger, 1986; Zeger & Liang, 1986). The GEE algorithm is available in most statistics programs (SPSS, SAS, STATA), and Ballinger (2004) provides an excellent description of how to analyze data using GEE. When one is interested in modeling patterns of change over time with continuous measures of behaviors, growth curve models can be estimated using standard multilevel modeling programs, such as the MIXED procedure in SPSS (Proc Mixed in SAS).

As discussed in the section on behavioral observations with experimental manipulations, in some cases, behaviors are measured within dyadic contexts, such as during interactions between romantic partners or between two newly acquainted partners. When both partners provide behavioral data, their behaviors are likely nonindependent (e.g., how one romantic partner behaves is likely correlated with how her partner behaves within the interaction). In this handbook, Kenny and Kashy provide an overview of how to analyze dyadic data (see also Kenny & Kashy, 2011; Kenny, Kashy, and Cook, 2006). The same basic principles described in these papers apply to analyzing behavioral data that are dyadic in nature .

Sequential Analysis

Although what people do when interacting is important, how interactions unfold across time is possibly more important. With many phenomena, from the courting behavior of birds to the escalation of human conflict, the patterning of behavior is critical – “a defining characteristic of interaction is that it unfolds in time” (Bakeman & Gottman, 1997, p. 1). Furthermore, Gottman and Roy (1990, p. 1) contend that: “the dimension of time is so central to conceptualizing social interaction that its use will lead us to think of interaction itself as temporal form.”

How did Gottman and colleagues come to conclude that sequence is a central (if not *the* central) issue in understanding behavior? First, Gottman and Roy (1990) discuss several research instances – family management, couples interaction, and schoolchildren's peer interactions – in which base rate analyses show no difference between functional and dysfunctional groups, but analyses of sequence show strong differences between groups. Second, sequential analyses sometimes reveal unexpected patterns and thus serve as a theory-generating, as

well as a theory-testing, tool.

Unidirectional dependence. Most studies that have used sequential analysis have tested if one person's behavior follows the other's behavior at a rate higher than chance. For example, does one roommate's blame follow the other's blame more than what would be expected by chance? This is a one-way, or unidirectional, test of linkage between the two behaviors.

There are two forms of significant linkage between behaviors. First, compared with chance, the antecedent behavior can increase the likelihood of the consequent behavior. This is an escalation effect. Second, the antecedent behavior can decrease the likelihood of the consequent behavior. This is a suppression effect.

Bidirectional dependence. Bidirectional dependence simultaneously tests if A results in B and if B results in A. For example, we could simultaneously test if roommate A's blame follows roommate B's and if B's blame follows A's blame (i.e., reciprocity). The same logic for escalation and suppression effects applies to bidirectional dependence tests. Wampold (1989) provides a formula for conducting a bidirectional test.

Although one could perform two unidirectional tests, the bidirectional test is superior for three reasons (Wampoldt & Margolin, 1982). First, if one is interested in reciprocity by both partners, the bidirectional test is more appropriate. Second, because unidirectional tests are not independent, multiple tests result in either inflation of the alpha level or a decrease in power owing to the use of the Bonferroni inequality. Third, it is possible for the bidirectional test to be significant, even when each of the unidirectional tests is not.

Dominance. Often researchers are interested in who is the more dominant person in an interaction. Gottman and Ringland (1981, p. 395) defined dominance as an “asymmetry in predictability; that is, if B's behavior is more predictable from A's past [behavior] than conversely, A is said to be dominant.” Thus, dominance indicates who is leading the dance. (Note that the label is referring to statistical dominance, which is not necessarily the same as perceived dominance or behavior that might be labeled as domineering.) There are two forms of dominance. In parallel dominance, the same two behaviors are considered. For example, is a student's hostility more predictable following the teacher's hostility than the other way around ?

Structure of data. To analyze data sequentially, three conditions must be met (Bakeman & Gottman, 1997). First, the temporal sequence must be preserved.

Thus, tallies of frequencies (i.e., base rates) are not sufficient; the coding must reflect the order in which the behaviors were performed. Second, codes must be mutually exclusive (i.e., only one code per event). Third, the codes must be exhaustive (i.e., there is a code for each behavior). To construct the matrices needed to test such patterning, it is useful to think of a moving window that can slide over the data stream (only data within the parentheses are visible). Working with a window the size of two events, the analysis would proceed until all the pairs are accounted for. A *transition matrix* (see [Table 14.5](#)) contains the tally of the pairs revealed by the moving window. If the events are not contiguous, a transition matrix for a specified lag can be computed. The rows specify the lag 0 (i.e., present) behaviors of a wife, and the columns specify the lag 1 (i.e., immediate past) behaviors of her husband. The frequencies represent the number of times that each lag 0 wife behavior is preceded by a husband behavior at lag 1. The example data suggest that husband behavior tends to be reciprocated with like behavior of the wife (e.g., positives are mostly met with positives).

Table 14.5. Contingency Table (Transition Matrix) of Lagged Effects of Dyadic Behavior

Lag 0 Partner 1	Lag 1 Partner 2 Behavior		
Behavior	Positive	Neutral	Negative
Positive	20	10	0
Neutral	10	20	10
Negative	0	10	20

Once the transition matrix is formed, the conditional probability (i.e., the probability of a behavior being emitted, given a particular antecedent behavior) can be computed. If a conditional probability is, say, 0.75, does that constitute an important pattern? This is not known until we know if the conditional probability exceeds chance.

Thus, the null hypothesis in sequential analysis states that the behaviors are randomly ordered and that any apparent patterns are attributable to chance. A z-

score – derived by Sackett (1979) and later modified by Allison and Liker (1982) and Wampold and Margolin (1982) – can be computed to test for the deviation from chance. However, despite their widespread use, z-scores have a major Achilles heel – they are influenced by the length (total number of transitions in the interaction) and by the base rates of the two behaviors under examination. Thus, the same degree of contingency will produce different z-scores across different dyads because of these factors. Nonparametric statistics such as Yule's Q (Bakeman & Quera, 2011) and Wampold Kappa (Wampold, 1989) have been offered and we advise their use. The sequential variables can then be used as scores in calculating test statistics (e.g., correlations, structural equation modeling, multilevel modeling) .

Loglinear Approach to Sequential Analysis

Loglinear methods (Bakeman & Quera, 1995) provide a flexible alternative approach to sequential analysis. The loglinear approach begins with the multidimensional contingency table consisting of frequencies of given behaviors and compares the fit of the observed data to patterns that would result given (the lack of) researcher-specified patterns of association. The simplest version is given by the two-way table, an example of which is found in Table 14.5.

Traditional sequential approaches would model each of the contingencies from Table 14.5 individually, for example, forming Allison and Liker z-scores representing positive→positive, neutral→neutral, and negative→negative contingencies. Such contingencies can be estimated in the loglinear context as well, although via the likelihood ratio chi-square (G^2). However, the loglinear approach is more flexible in that all three contingencies can be tested at once, similar to an omnibus ANOVA testing differences among three groups' means in a single test, thus protecting against Type-I error (Bakeman & Quera, 1995). Loglinear analysis follows the traditional rationale of the chi-square test of independence in which the observed frequencies are compared with expected frequencies that would be obtained if there were no association (a “no two-way interaction model,” in loglinear terms). A significant G^2 indicates the presence of a significant lagged association.

The above is a simple example of a loglinear approach to sequential analysis. However, the flexibility of the loglinear approach is that it may be generalized beyond the two-way case to accommodate higher order interactions (e.g., three-way, four-way). For example, one might hypothesize that lag 1 negative husband behavior is less likely reciprocated by the wife at lag 0 if the husband was

positive or neutral at lag 2. These frequencies would be represented and tested in a three-way contingency table: lag 0 wife behavior \times lag 1 husband behavior \times lag 2 husband behavior.

Examples with sophisticated applications of the loglinear approach are found in the study of in attorney-witness exchanges in the courtroom (Gnisci & Bakeman, 2007) and couples interaction (Notarius et al., 1989). Loglinear analysis can be carried out in standard statistical programs such as SPSS. Bakeman also offers a stand-alone program called ILOG (see Bakeman & Robinson, 1994) at his website: www2.gsu.edu/~psyraab/BakemanPrograms.htm. Furthermore, loglinear analysis can be performed in Mplus (Muthén & Muthén, 2010), as described in the “Multilevel Loglinear Analysis” section later in the chapter.

Dimensional Analyses of Behavior Sequences

The term “sequential analysis” is usually applied to patterns among categorical observations over the course of an interaction. In many instances, however, researchers are interested in patterns among dimensionally measured behaviors, such as whether the intensity of one person's behavior depends on the intensity of another person's prior behavior. Some examples of this are found in studies of the synchrony of parent-child and adult-adult interaction (e.g., Feldman, 2007; Dowdney & Pickles, 1991; Julien, Brault, Chartrand, & Bégin, 2000; Warner, 1992) and the coordination within and between people of emotional behavior, emotion experience, and physiology (e.g., Butler, 2011; Guastello, Pincus, & Gunderson, 2006; Levenson & Gottman, 1983; Mauss, Levenson, McCarter, Wilhelm, & Gross, 2005). Such processes usually require a different set of statistics than is used in traditional sequential analysis.

Broadly, dimensional approaches to behavior sequences are usually aimed at establishing the influence of one person's behavior on another person's behavior, the mutual coordination of behaviors in dyads, or sometimes the uninfluenced aspects of social interactions. They make use of what is referred to as *time series* data. A behavioral time series consists of observations made repeatedly at regular intervals of time, such as the intensity of positive emotion rated for each consecutive five-second interval over the course of an interaction. Patterns in observational time series data can be studied in either the time domain or the frequency domain. The time domain approach is far more common and will thus be emphasized here. Readers interested in the frequency domain approach are referred to Warner (1992) and Richardson, Dale, & Marsh, Chapter 11 in this

volume.

Time domain approach. Analyses in the time domain are cast in the familiar terms of correlation and regression, with the primary difference being that the correlations are within-subject or within-dyad. The researcher looks for evidence that present behaviors are correlated between interactive partners and/or that present behaviors are correlated with a person's own past behavior or the past behavior of the partner. The prevailing statistical techniques are cross-correlation and time series regression, although see the “Dynamical Systems Modeling” section later in the chapter for an alternative approach.

Cross-correlation analysis simply looks for the correlations between one person's behavior in the present (lag 0) with the behavior of another person at various lags (i.e., points in the past). It is up to the investigator to determine which of the various possible cross-correlations s/he is interested in. There may, for example, be substantive reasons for a focus on relatively short-term or long-term effects. Extensive data preparation is required prior to cross-correlation analysis. Each time series must be “prewhitened,” which means that any cycles and trends over time must be removed.

Simply establishing the extent of cross-correlation between behaviors may be of substantive interest. However researchers are often interested in modeling differences in the degree of cross-correlation *among dyads* in relation to other variables. In these cases, the cross-correlation calculated within each dyad with time series methods is used as a variable in other analyses (e.g., Pearson correlation). An example of this is found in Feldman's work in which greater mother-infant synchrony (i.e., greater cross-correlations of mother and infant behavior) was associated with better child outcomes (e.g., Feldman, 2007).

Time series regression takes a similar approach to cross-correlational analysis, with two key differences. First, autoregressive effects (i.e., the internal predictability of a person's behavior across time) are part of the model estimation rather than a prior step; such effects are removed when prewhitening variables prior to cross-correlational analysis. Second, because lagged terms are simultaneously entered, each lagged term represents a *unique* association of past and present behavior, controlling for all other lagged effects in the model.

Warner (1992) describes an approach to time series regression for situations in which the researcher wishes to model both the internal (i.e., autoregressive) and social determinants (i.e., partner effects) in behavior times series. The internal and social determinant estimates (R^2) within each dyad may be of substantive

interest. However, frequently, investigators are interested in differences in these parameters from dyad to dyad. In such cases, the R^2 , like the cross-correlation, can be treated as variables for subsequent analysis. We have used this analytic method to model the impact of child behavior on maternal emotion and how the degree of child influence and the degree of autocorrelation predict maternal discipline practices (Lorber & Slep, 2005).

Time series regression and cross-correlation are available in SPSS Trends (which can be purchased as an add-on) and also in the freely downloadable program, R (R Development Core Team, 2005). R offers a much greater array of time series analytic models and has the added advantage of several automated model selection packages for the prewhitening of data .

Recent Developments in Analyzing Observational Data

In this section we briefly describe recent analytic developments for observational data that will likely be of interest to many social-psychological researchers. The list is certainly not comprehensive. Instead, the featured analytic models were selected with an eye toward relevance to social-psychological applications and feasibility of implementation (e.g., availability of computer programs).

Dynamical Systems Modeling. In the mid-1990s, Gottman, Murray, and their colleagues developed a set of nonlinear difference equations – a “dynamical system” – to model change over time in couples’ behavior (e.g., Cook et al., 1995). These methods were probably opaque to many researchers and were not implemented in common software packages; they appear to have been used primarily by their progenitors (e.g., Gottman et al., 2003; Gottman, Ryan, Swanson, & Swanson, 2005). However, recent work by Hamaker and colleagues has set the stage for more widespread usage (Hamaker, 2009; Hamaker, Zhang, & Van der Mass, 2009; Madhyastha, Hamaker, & Gottman, 2011). Richardson, Dale, and Marsh (Chapter 11 in this volume) provide a broad overview to dynamical systems models in social psychology.

Briefly, this collection of techniques analyzes dimensional time series data from dyads. The techniques are designed to capture uninfluenced steady states as well as multiple types of nonlinear influence from one person to another. Uninfluenced steady states refer to what each person “brings to the table” – for example, one’s overarching emotional style. To illustrate two of the many possibilities for influence: (1) negative behavior in one spouse might have a

linear association with the degree of subsequent partner negativity, with rises and falls in one person's behaviors predicting similar rises and falls in the other's behavior, and (2) there could be thresholds above and below which the relation of spousal negativity and subsequent negativity in the partner changes (e.g., a person may “ignore” low-level partner negativity, respond in kind to moderate partner negativity, and withdraw from high-level partner negativity). The innovation of Hamaker *et al.* (2009) was the realization that Gottman and Murray's models were special cases of the previously established threshold autoregressive model. The advantage is that there are preestablished influence functions that can be evaluated, methods of parameter estimation, and statistical criteria for selecting from among the different influence models (via the common BIC statistic).

Research using these tools is in its infancy. For example, Madhyastha *et al.* (2011) showed that many couples do not exhibit reliable interpartner influence (i.e., uninfluenced steady states were very powerful), and that partners from different couples differ in how they influence each other. Given the availability of the “dyad” statistical package for R (Madhyastha & Hamaker, 2009; R Development Core Team, 2005), a free and very powerful statistical program, there is great untapped potential for the application of nonlinear dynamical modeling in the context of threshold autoregressive models. Such models would be of interest in any social psychological research in which social influence in dimensionally rated behavior dyads might be expected to be nonlinear, and/or in which the estimation of what each dyad member contributes to social interaction, independent of her/his partner's behavior, is of interest.

Multilevel Survival Analysis. Stoolmiller and Snyder (2006; Snyder, Stoolmiller, Wilson, & Yamamoto, 2003) offer a novel approach to characterizing the course of an interaction, utilizing a variant of survival analysis for repeated events based on prior work (Gardner & Griffin, 1989; Griffin & Gardner, 1989). In traditional survival analysis, time to a single event is modeled as the function of predictors or covariates. For example, if one were interested in gender differences in longevity, time to death would be modeled as a function of gender. In contrast, the events of interest in behavioral observation are most often free to repeat. Thus, in the present context, survival analysis is adapted to model repeated events within dyads. The “hazard rate” – time between displays of a behavior – is the focus of these analyses. It can be modeled as a function of static or unchanging covariates (e.g., gender) and time-varying covariates (e.g., behavior of another person and experimental manipulations) via Cox regression, which is one type of a survival model.

Hazard rates and their associations with covariates usually vary from dyad to dyad, giving rise to the need to model them in a multilevel framework (Raudenbush & Bryk, 2001; Schoemann, Rhemtulla, & Little, Chapter 21 in this volume). Snyder *et al.* (2003) provide an example of how multilevel survival analyses can be used to model emotional displays in dyadic interaction. The time between displays of child anger in parent-child interactions (i.e., hazard rate) was modeled as a function of several static and time-varying covariates. The authors found that the time between children's anger displays decreased over the course of interactions with their parents the more the parents' insensitive and negative behaviors toward the child accrued, illustrating a time-varying covariate effect within dyads. Moreover, children who were rated by their parents as more antisocial (i.e., aggressive and oppositional) had decreased time between anger displays, illustrating a static covariate effect. However, the authors found no evidence that the dynamic link between parenting and child anger was related to parent-or teacher-reported antisocial child behavior, illustrating how one might test the association of a static, between-dyad covariate (e.g., score on a questionnaire) with a within-dyad dynamic pattern of observed behavior over the course of social interaction.

The multilevel survival approach has, to our knowledge, not yet been employed in social psychology. Nonetheless, Butler (2011) recently pointed out the broad relevance of this approach to what she terms “temporal interpersonal emotion systems” in social interaction, for modeling emotion reciprocity, reactivity, and escalation and de-escalation. The multilevel survival approach has applicability in any setting in which the time between a behavior's occurrence, whether an emotion display or some other behavior, marks a process of theoretical interest.

At present, multilevel survival analyses are available in S-Plus (Insightful Corp., 2001), with S-Plus survival analysis code and SPSS data preparation syntax available from Stoolmiller and Snyder online at [dx.doi.org/10.1037/1082-989X.11.2.164.supp](https://doi.org/10.1037/1082-989X.11.2.164.supp).

Multilevel Loglinear Analysis. Dagne and Howe (Dagne *et al.* 2002; Howe *et al.*, 2005) recently developed a multilevel extension of loglinear analysis for observational data (see “Loglinear Approach to Sequential Analysis” section earlier in the chapter). This method has multiple advantages over traditional loglinear analyses of behavior observations. To name a few, it has superior handling of the nesting of behavior inherent in many studies of social behavior, where behaviors are often nested within episodes (e.g., experimental conditions)

that are further nested within dyads. Multilevel loglinear analysis further deftly handles cases with low rates of target behaviors, a common problem in observational research as estimates of sequential patterns among low-rate behaviors have greater measurement error; such cases are weighted to a lesser extent than are higher-rate cases. Moreover, it provides an analytic framework to model sequential patterns as a function of other variables.

Howe *et al.* (2005) use the example of behavior in married couples to illustrate the techniques. Sequences of interest are first estimated within dyads, for example, husband reciprocation of wife negativity. Because these patterns occur at different rates in different couples, they are modeled as random effects (e.g., the degree of negative reciprocity is allowed to freely vary among couples). The sample wide average of each random effect (e.g., the overall strength of negative reciprocity) can be compared against zero to test for a significant overall sequential association. The random effects or sequences can then be modeled in relation to other random effects, answering such questions as whether couples who reciprocate one another's positive behaviors at a high rate are less likely to reciprocate negative behaviors. Behavior sequences or random effects can also be modeled in relation to other consequential variables such as experimental manipulations (e.g., conflict vs. events of the day discussions) and individual or dyad level characteristics (e.g., personality and relationship adjustment). Finally, contrasts can also be structured to compare the relative strength of different sequences (e.g., whether men are more likely than women to reciprocate negative behavior).

The multilevel loglinear approach has, to our knowledge, not yet been employed in social psychology. However, it is a very flexible approach, with wide applicability to questions of interest of social psychologists who seek to understand behavioral sequences in dyads. Moreover, it is clearly superior to the ordinary loglinear approach to sequential analysis. Howe *et al.* (2005) offer example syntax for implementing multilevel loglinear analysis in Mplus (Muthén & Muthén, 2010). Moreover, sample Mplus data, input, and output files corresponding to the examples in Dagne *et al.* (2002) are provided at statmodel.com.

Conclusions and Future Directions

We began this chapter noting that there is nothing as practical as a good theory-testing tool. Behavioral observation is a research method centered on the

identification of behaviors worth theorizing and enables the testing of theories of behavior. The video and audio records that are created in many behavioral observation studies provide one of the richest sources of information available in the social sciences for the study of social interactions. Because videos (unlike live observations) are archivable, they can be used in the future to investigate different hypotheses or as improved methods are developed. Behavioral observation serves as a compliment to a wide variety of self-report methods on behavior, cognition, and affect, as well as an intriguing partner to studies that employ other data collection methods, including biological assays. In short, behavioral observation has great potential for advancing knowledge in social and personality psychology.

Over the past 50 years, a number of significant advances have occurred in behavioral observation methodology, most notably in terms of the complexity of coding systems, the recording of observations and of codes, and statistical analyses. Each of these advances has been related to advances in computer technology. At present, the field is poised for an explosion in opportunities. With the penetration of smart phones and other digital technology, never has recording been so easy, cheap, and ubiquitous or the opportunities for observation been so plentiful (Mehl & Conner, 2012). Natural sampling methods (like the EAR) will become easier and easier. Further, computerized coding of behavior without the need for human coders (e.g., Black et al., in press; Cohn, Zlochower, Lien, & Kanade, 1999) already exists and will likely increase in its availability and impact in the coming decade.

What has been lacking, however, is the accumulation of information on the reliability and validity of coding systems, beyond the focus of the field on interrater agreement. Certainly, interrater agreement is vital to the value of a coding system, but it is not the sole issue of interest. At this point, there are a number of general-purpose coding systems that have been used by multiple researchers with varied interests over time, and further attention to the basic properties of these systems is needed. Unfortunately, finding funding to do this type of work is difficult and “nuts and bolts” research is not flashy. With a renewed focus in this area, and the continued innovation that has been at the core of the method throughout its short history, behavioral observation is well positioned to push the field of social and personality psychology forward in the coming years.

Over the past 20 years, the study of behavior in social psychology has rapidly declined, with the majority of studies collecting self-reported measures (e.g.,

paper and pencil ratings; Baumeister, Vohns, & Funder, 2007), with some notable exceptions (e.g., behavioral measures of implicit attitudes; see Gawronski and De Houwer, [Chapter 12](#) in this volume, for a review). Scholars may be discouraged from collecting behavioral data in part because of the complexities involved in collecting, coding, and analyzing it. Behavior is also hard to change, and designing an experimental manipulation that alters “actual behavior” may prove daunting for many researchers. For these reasons (and more), researchers may be discouraged from collecting behavioral data. However, behavioral data can provide insight into psychological processes that other dependent measures cannot do alone – it is what put social psychology “on the map” many decades ago and it is at the heart of many of our theories. New scholars should be encouraged to know that there are many basic questions that still remain unanswered in social psychology that can only be answered with behavioral data and that, although not without its challenges, collecting behavioral data is certainly worth the effort.

References

- Allison, P. D., & Liker, J. K. (1982). Analyzing sequential categorical data on dyadic interaction: Comment on Gottman. *Psychological Bulletin*, 91, 393–403.
- Altman, I., & Taylor, D. A. (1973). *Social penetration: The development of interpersonal relationships*. New York: Holt, Rinehart, & Winston.
- Aron, A., Norman, C., Aron, E., McKenna, C., & Heyman, R. E. (2000). Couples shared anticipation in novel and arousing activities and experienced relationship quality. *Journal of Personality and Social Psychology*, 78, 273–284.
- Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Groups, Leadership, and Men*, 5, 222–236.
- Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). New York: Cambridge University Press.
- Bakeman, R., McArthur, D., Quera, V., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, 2, 357–370.
- Bakeman, R., & Quera, V. (1995). Loglinear approaches to lag-sequential

- analysis when consecutive codes may and cannot repeat. *Psychological Bulletin*, 118, 272–284.
- Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. New York: Cambridge University Press.
- Bakeman, R., & Robinson, B. F. (1994). *Understanding loglinear analysis with ILOG: An interactive approach*. Hillsdale, NJ: Erlbaum.
- Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods*, 7, 127–150.
- Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, 2, 396–403.
- Black, M. P., Katsamanis, A., Baucom, B. R., Lee, C. C., Lammert, A. C., Christensen, A., Georgiou, P. G., & Narayanan, S. S. (2013). Toward automating a human behavioral coding system for married couples interactions using speech acoustic features. *Speech Communication*, 55(1), 1–21.
- Blascovich, J., Mendes, W. B., Hunter, S. B., Lickel, B., & Kowai-Bell, N. (2001). Perceiver threat in social interactions with stigmatized others. *Journal of Personality and Social Psychology*, 80, 253–267.
- Butler, E. A. (2011). Temporal interpersonal emotion systems: The “TIES” that form relationships. *Personality and Social Psychology Review*, 15, 367–393.
- Campos, B., Graesch, A. P., Repetti, R., Bradbury, T., & Ochs, E. (2009). Opportunity for interaction? A naturalistic observation study of dual-earner families after work and school. *Journal of Family Psychology*, 23, 798–807.
- Christensen, A., & Hazzard, A. (1983). Reactive effects during naturalistic observation of families. *Behavioral Assessment*, 5, 349–362.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6, 284–290.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for

- scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213.
- Cohn, J. F., Zlochower, A. J. Lien, J., & Kanade, T. (1999). Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. *Psychophysiology*, 36, 35–43.
- Cook, J., Tyson, R., White, J., Rushe, R., Gottman, J., & Murray, J. (1995). Mathematics of marital conflict: Qualitative dynamic mathematical modeling of marital interaction. *Journal of Family Psychology*, 9, 110–130.
- Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Dadds, M. R., & McHugh, T. A. (1992). Social support and treatment outcome in behavioral family therapy for child conduct problems. *Journal of Consulting and Clinical Psychology*, 60, 252–259.
- Dagne, G., Howe, G. W., Brown, C. H., & Muthén, B. (2002). Hierarchical modeling of sequential behavioral data: An empirical Bayesian approach. *Psychological Methods*, 7, 262–280.
- Dowdney, L., & Pickles, A. R. (1991). Expression of negative affect with disciplinary encounters: Is there dyadic reciprocity? *Developmental Psychology*, 27, 606–617.
- Eid, M., & Diener, E. (Eds.) (2006). *Handbook of multimethod measurement in psychology*. Washington, DC: American Psychological Association.
- Feldman, R. (2007). Parent-infant synchrony and the construction of shared timing: Physiological precursors, developmental outcomes, and risk conditions. *Journal of Child Psychology and Psychiatry*, 48, 329–354.
- Festinger, L., Riecken, H., & Schacter, S. (1956). *When prophecy fails: A social and psychological study of a modern group that predicted the destruction of the world*. New York: Harper and Row.
- Foster, D. A., Caplan, R. D., & Howe, G. W. (1997). Representativeness of

observed couple interaction: Couples can tell, and it does make a difference. *Psychological Assessment*, 9, 285–294.

Gardner, W., & Griffin, W. A. (1989). Methods for the analysis of parallel streams of continuously recorded social behaviors. *Psychological Bulletin*, 105, 446–455.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. New York: W. H. Freeman.

Gnisci, A., & Bakeman, R. (2007). Sequential accommodation of turn taking and turn length: A study of courtroom interaction. *Journal of Language and Social Psychology*, 26, 234–259.

Goff, P. A., Steele, C. M., & Davies, P. G. (2008). The space between us: Stereotype threat and distance in interracial contexts. *Journal of Personality and Social Psychology*, 94, 91–107.

Goodenough, F. L. (1931). *Anger in young children*. Minneapolis: University of Minnesota Press.

Gottman, J., Ryan, K., Swanson, C., Swanson, K. (2005). Proximal change experiments with couples: A methodology for empirically building a science of effective interventions for changing couples' interaction. *Journal of Family Communication*, 5, 163–190.

Gottman, J. M. (1978). Nonsequential data analysis techniques in observational research. In G. P. Sackett (Ed.), *Observing behavior: Vol. 2. Data collection and analysis methods* (pp. 45–61). Baltimore: University Park Press.

Gottman, J. M. (1979). *Marital interaction*. Champaign, IL: Research Press.

Gottman, J. M. (1994). *What predicts divorce? The measures*. Hillsdale, NJ: Erlbaum.

Gottman, J. M. (1999). *The marriage clinic: A scientifically-based marital therapy*. New York: Norton.

Gottman, J. M., & Krokoff, L. J. (1989). Marital interaction and satisfaction: A longitudinal view. *Journal of Consulting and Clinical Psychology*, 57, 47–52.

Gottman, J. M., Levenson, R. W., Swanson, C., Swanson, K., Tyson, R., & Yoshimoto, D. (2003). Observing gay, lesbian and heterosexual couples' relationships. *Journal of Homosexuality*, 45, 65–91.

- Gottman, J. M., & Ringland, J. T. (1981). The analysis of dominance and bidirectionality in social development. *Child Development*, 393–412.
- Gottman, J. M., & Roy, A. K. (1990). *Sequential analysis: A guide for behavioral researchers*. Cambridge: Cambridge University Press.
- Gray, H. M., Mendes, W. B., Denny-Brown, C. (2008). An in-group advantage in detecting intergroup anxiety. *Psychological Science*, 19, 1233–1237.
- Greenwald, A. G., Nosek, B., & Banaji, M. R. (2003). “Understanding and using the Implicit Association Test: I. An improved scoring algorithm”: Correction to Greenwald *et al.* (2003). *Journal of Personality & Social Psychology*, 85, 197–216.
- Griffin, W. A. (2000). A conceptual and graphical method for converging multisubject behavioral observational data into a single process indicator. *Behavior Research Methods, Instruments, and Computers*, 32, 120–133.
- Griffin, W. A., & Gardner, W. (1989). Analysis of behavioral durations in observational studies of social interaction. *Psychological Bulletin*, 106, 497–502.
- Guastello, S. J., Pincus, D., & Gunderson, P. R. (2006). Electrodermal arousal between participants in a conversation: Nonlinear dynamics and linkage effects. *Nonlinear Dynamics, Psychology, and Life Sciences*, 10, 365–399.
- Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Interrater Reliability*, 1, 1–5.
- Gwet, K. (2008). Variance estimation of nominal-scale interrater reliability with random selection of raters. *Psychometrika*, 73, 407–430.
- Hamaker, E. L. (2009). Determining the number of regimes in threshold autoregressive models by means of information criteria. *Journal of Mathematical Psychology*, 53, 518–529.
- Hamaker, E. L., Zhang, Z., & Van der Maas, H. L. J. (2009). Using threshold autoregressive models to study dyadic interactions. *Psychometrika*, 74, 727–745.
- Haney, C., Banks, W., & Zimbardo, P. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology*, 1, 69–

97.

- Hawes, D. J., & Dawes, M. R. (2006). Assessing parenting practices through parent-report and direct observation during parent-training. *Journal of Child and Family Studies*, 15(5), 555–568.
- Haynes, S. N., & O'Brien, W. H. (2000). *Principles and practice of behavioral assessment*. New York: Kluwer.
- Heyman, R. E. (2001). Observation of couple conflicts: Clinical assessment applications, stubborn truths, and shaky foundations. *Psychological Assessment*, 13, 5–35.
- Heyman, R. E. (2004). Rapid Marital Interaction Coding System. In P. K. Kerig & D. H. Baucom (Eds.) *Couple observational coding systems* (pp. 67–94). Mahwah, NJ: Lawrence Erlbaum Associates.
- Heyman, R. E., Chaudhry, B. R., Treboux, D., Crowell, J., Lord, C., Vivian, D., & Waters, E. B. (2001). How much observational data is enough? An empirical test using marital interaction coding. *Behavior Therapy*, 32, 107–123.
- Heyman, R. E., Eddy, J. M., Weiss, R. L., & Vivian, D. (1995). Factor analysis of the Marital Interaction Coding System. *Journal of Family Psychology*, 9, 209–215.
- Holley, W., & Guilford, J. P. (1964). A note on the G-index of agreement. *Educational and Psychological Measurement*, 24, 749–754.
- Hops, H., Davis, B., & Longoria, N. (1995). Methodological issues in direct observation: Illustrations with the Living in Familial Environments (LIFE) coding system. *Journal of Clinical Child Psychology*, 24, 193–203.
- Howe, G. W., Dagne, G., & Brown, C. H. (2005). Multilevel methods for modeling observed sequences of family interaction. *Journal of Family Psychology*, 19, 72–85.
- Insightful Corp. (2001). *S-Plus 6 for Windows user's guide*. Seattle: Author.
- Jacob, T., Tennenbaum, D., Seilhamer, R. A., Bargiel, K., & Sharon, T. (1994). Reactivity effects during naturalistic observation of distressed and nondistressed families. *Journal of Family Psychology*, 8, 354–363.
- Jacobson, N. S. (1985). The role of observational measures in behavior therapy

- outcome research. *Behavioral Assessment*, 7, 297–308.
- Julien, D., Brault, M., Chartrand, E., & Begin, J. (2000). Immediacy behaviors and synchrony in satisfied and dissatisfied couples. *Canadian Journal of Behavioural Science*, 32, 84–90.
- Kelly, M. B. (1977). A review of the observational data-collection and reliability procedures reported. *Journal of Applied Behavior Analysis*, 10, 97–101.
- Kenny, D. A., & Kashy, D. A. (2011). Dyadic data analysis using multilevel modeling. In J. J. Hox & J. K. Roberts (Eds.), *Handbook for advanced multilevel analysis* (pp. 335–370). New York: Routledge/Taylor & Francis Group.
- Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York: Guilford.
- Kenny, D. A., Mohr, C. D., & Levesque, M. J. (2001). A social relations variance partitioning of dyadic behavior. *Psychological Bulletin*, 127, 128–141.
- Kerig, P. K., & Baucom, D. H. (Eds.). (2004). *Couple observational coding systems*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Kerig, P. K., & Lindahl, K. M. (Eds.). (2000). *Family observational coding systems*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Krokoff, L. J., Gottman, J. M., & Hass, S. D. (1989). Validation of a global rapid couples interaction scoring system. *Behavioral Assessment*, 11, 65–79.
- Lakin, J. L., Chartrand, T. L., & Arkin, R. M. (2008). I am too just like you: Nonconscious mimicry as an automatic behavioral response to social exclusion. *Psychological Science*, 19, 816–822.
- Lewin, K. (1951). *Field theory in social science*. Chicago: University of Chicago Press.
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.
- Lorber, M. F. (2006). Can minimally trained observers provide valid global ratings? *Journal of Family Psychology*, 20, 335–338.
- Lorber, M. F., & Slep, A. M. S. (2005). Mothers' emotion dynamics and their

- relations with harsh and lax discipline: microsocial time series analyses. *Journal of Clinical Child and Adolescent Psychology*, 34, 559–568.
- Lorenz, K. (1970). *Studies in animal and human behaviour*, Vol. 1 (R. Martin, Transl.). Cambridge, MA: Harvard University Press.
- Lorenz, K. (1971). *Studies in animal and human behaviour*, Vol. 2 (R. Martin, Transl.). Cambridge, MA: Harvard University Press.
- Madhyastha, T., & Hamaker, E. (2009). *Dyad*. Retrieved from <http://cran.r-project.org/web/packages/dyad/index.html>
- Madhyastha, T. M., Hamaker, E. L., & Gottman, J. M. (2011). Investigating spousal influence using moment-to-moment affect data from marital conflict. *Journal of Family Psychology*, 25, 292–300.
- Margolin, G., Burman, B., & John, R. (1989). Home observations of married couples reenacting naturalistic conflicts. *Behavioral Assessment*, 11, 101–118.
- Margolin, G., Oliver, P., Gordis, E., O’Hearn, H. G., Medina, A. M., Ghosh, C. M., & Morland, L. (1998). The nuts and bolts of behavioral observation of marital and family interaction. *Clinical Child and Family Psychology Review*, 1, 195–213.
- Martinez Jr., C. R., & Forgatch, M. S. (2001). Preventing problems with boys’ noncompliance: Effects of a parent training intervention for divorcing mothers. *Journal of Consulting and Clinical Psychology*, 69, 416.
- Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, 5, 175–190.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30–46.
- Mead, M. (1928) *Coming of age in Samoa: A psychological study of primitive youth for Western civilisation*. New York: William Morrow & Co.
- Mehl, M. R. (2007). Eavesdropping on health: A naturalistic observation approach for social health research. *Social and Personality Psychology Compass*, 1, 359–380.
- Mehl, M. R. & Conner, T. S. (Eds.). (2012). *Handbook of research methods for studying daily life*. New York: Guilford Press.

- Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90, 862–877.
- Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, 84, 857–870.
- Mehl, M. R., Pennebaker, J. W., Crow, M. D., Dabbs, J., & Price, J. H. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, and Computers*, 33, 517–523.
- Mehl, M. R. & Robbins, M. L. (2012). Naturalistic observation sampling: The Electronically Activated Recorder (EAR). In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 176–192). New York: Guilford Press.
- Mehl, M. R. Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. (2007). Are women really more talkative than men? *Science*, 317, 82.
- Mendes, W. B., Major, B., McCoy, S., & Blascovich, J. (2008). How attributional ambiguity shapes physiological and emotional responses to social rejection and acceptance. *Journal of Personality and Social Psychology*, 94, 278.
- Mitchell, S. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, 86, 376–390.
- Muthén, L. K., & Muthén, B. O. (2010). *Mplus User's Guide* (6th ed). Los Angeles, CA: Author.
- Notarius, C. I., Benson, P. R., Sloane, D., Vanzetti, N. A. *et al.* (1989). Exploring the interface between perception and behavior: An analysis of marital interaction in distressed and nondistressed couples. *Behavioral Assessment*, 11, 39–64.
- Pasch, L. A., & Bradbury, T. N. (1998). Social support, conflict, and the development of marital dysfunction. *Journal of Consulting and Clinical Psychology*, 66, 219–230.
- Patterson, G. R. (1982). *Coercive family process*. Eugene, OR: Castalia.

- Patterson, G. R., Reid, J. B., & Dishion, T. J. (1992). *Antisocial boys*. Eugene, OR: Castalia.
- Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89, 277–293.
- Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. (2005). Prosocial behavior: Multilevel perspectives. *Annual Review of Psychology*, 56, 365–392.
- Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, 112, 160–164.
- Quera, V., Bakeman, R., & Gnisci, A. (2007). Observer agreement for event sequences: Methods and software for sequence alignment and reliability estimates. *Behavior Research Methods*, 39, 39–49.
- R. Development Core Team (2005). R: A language and environment for statistical computing, reference index version 2.2.1. R Foundation for Statistical Computing, Vienna, Austria.
- Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
- Reid, J. B. (Ed.) (1978). *A social learning approach, Vol. 2: Observation in home settings*. Eugene, OR: Castalia.
- Reid, J. B., Patterson, G. R., & Snyder, J. J. (2002). *Antisocial behavior in children and adolescents: A developmental analysis and the Oregon model for intervention*. Washington, DC: American Psychological Association Press.
- Robinson, B. F., & Bakeman, R. (1998). ComKappa: A Windows 95 program for calculating kappa and related statistics. *Behavior Research Methods, Instruments, and Computers*, 30, 731–732.
- Rusby, J., Estes, A., & Dishion, T. J. (1991). *Interpersonal process codes*. Unpublished coding manual, Oregon Social Learning Center, Eugene, OR.
- Sackett, G. P. (1979). The lag sequential analysis of contingency and cyclicity in behavioral interaction research. In J. D. Osofsky (Ed.), *Handbook of infant development* (pp. 623–649). New York: Wiley.
- Schmaling, K. B., Wamboldt, F., Telford, L., Newman, K. B., Hops, H., &

- Eddy, J. M. (1996). Interactions of asthmatics and their spouses: A preliminary study of individual differences. *Journal of Clinical Psychology in Medical Settings*, 3, 211–218.
- Shelton, J. N., & Richeson, J. (2006). Interracial interactions: A relational approach. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 38, pp. 121–181). San Diego, CA: Elsevier Academic Press.
- Shelton, K. K., Frick, P. J., & Wootton, J. M. (1996). Assessment of parenting practices in families of elementary school-aged children. *Journal of Clinical Child Psychology*, 25, 317–329.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.
- Shumway, R. H., & Stoffer, D. S. (2010). *Time series analysis and its applications: With R examples* (3rd ed.). New York: Springer.
- Smith, R. H., & Harris, M. J. (2006). Multimethod approaches in social psychology: Between-and within-method replication and multimethod assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 385–400). Washington, DC: American Psychological Association.
- Smoak, N. D., Scott-Sheldon, L. A. J., Johnson, B. T., & Carey, M. P. (2006). Sexual risk reduction interventions do not inadvertently increase the overall frequency of sexual behavior: A meta-analysis of 174 studies with 116,735 participants. *Journal of Acquired Immune Deficiency Syndromes*, 41, 374–384.
- Snyder, J., Edwards, P., McGraw, K., Kilgore, K., & Holton, A. (1994). Escalation and reinforcement in mother-child conflict: Social processes associated with the development of physical aggression. *Development and Psychopathology*, 6, 305–321.
- Snyder, J., Stoolmiller, M., Wilson, M., & Yamamoto, M. (2003). Child anger regulation, parental responses to children's anger displays, and early child antisocial behavior. *Social Development*, 12, 335–360.
- Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry*, 42, 725–728.
- Stoolmiller, M., & Snyder, J. (2006). Modeling heterogeneity in social

- interaction processes using multilevel survival analysis. *Psychological Methods*, 11, 164–177.
- Suen, H. K. (1988). Agreement, reliability, accuracy, and validity: Toward a clarification. *Behavioral Assessment*, 10, 343–366.
- Tashakkori, A., & Teddlie, C. (2010). *Handbook of mixed methods in social and behavioral research* (2nd ed.). Thousand Oaks, CA: Sage.
- Thornberry, T., & Brestan-Knight, E. (2011). Analyzing the utility of Dyadic Parent-Child Interaction Coding System (DPICS) warm-up segments. *Journal of Psychopathology and Behavioral Assessment*, 33, 187–195.
- Tickle-Degnen, L., & Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1, 285–293.
- Van Baaren, R. B., Janssen, L., Chartrand, T. L., & Dijksterhuis, A. (2009). Where is the love? The social aspects of mimicry. *Philosophical Transactions of the Royal Society B*, 1528, 2381–2389.
- Wampold, B. E. (1989). Kappa as a measure of pattern in sequential data. *Quality & Quantity*, 23, 171–187.
- Wampold, B. E., & Margolin, G. (1982). Non parametric strategies to test independence of behavioral states in sequential data. *Psychological Bulletin*, 92, 755–765.
- Wang, S. W., Repetti, R. L., & Campos, B. (2011). Job stress and family social behavior: The moderating role of neuroticism. *Journal of Occupational Health Psychology*, 16, 441–456.
- Warner, R. M. (1992). Sequential analysis of social interaction: Assessing internal versus social determinants of behavior. *Journal of Personality and Social Psychology*, 63, 51–60.
- Waters, E. B. (1978). The reliability and stability of individual differences in infant-mother attachment. *Child Development*, 49, 483–494.
- Whaley, S. E., Pinto, A., & Sigman, M. (1999). Characterizing interactions between anxious mothers and their children. *Journal of Consulting and Clinical Psychology*, 67, 826–836.
- Weiss, R. L., & Summers, K. J. (1983). *Marital Interaction Coding System III*. In E. E. Filsinger (Ed.), *A sourcebook of marriage and family assessment* (pp.

85–115). Beverly Hills, CA: Sage.

Whitehurst, G. J. (1984). Interrater agreement for journal manuscript reviews. *American Psychologist*, 39, 22–28.

Wieder, G. B., & Weiss, R. L. (1980). Generalizability theory and the coding of marital interactions. *Journal of Consulting and Clinical Psychology*, 48, 469–477.

Zeger, S. L., & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42, 121–130.

Chapter fifteen Methods for Studying Everyday Experience in Its Natural Context

Harry T. Reis, Shelly L. Gable and Michael R. Maniaci

Since the publication of the first edition of this *Handbook* in 2000, methods that go by the generic name of *everyday experience methods* have matured from the status of promising innovations to standard tools in social-personality psychology. A recent PsycInfo® search revealed more than 1,300 citations to studies conducted with one or another of these methods. Similarly, a 2009 survey indicated that 60% of editorial board members of social-personality journals had used these methods (Tracy, Robins, & Sherman, 2009). By *everyday experience methods* we refer not to a specific instrument or procedure but rather to a paradigm for studying social-psychological phenomena as they occur in the ebb and flow of everyday life – to “capture life as it is lived” (Bolger, Davis, & Rafaeli, 2003, p. 580). Everyday experience methods offer more than just another methodological alternative; their focus on ordinary, spontaneous activity allows researchers to evaluate theoretical models and hypotheses from a perspective that differs fundamentally from traditional methods. The payoff is a detailed, accurate, and multifaceted portrait of social behavior embedded in its natural context.

Under the heading of “everyday experience methods” we include diverse procedures and measures. Among the more popular examples are the experience sampling method (ESM; Hektner, Schmidt, & Csikszentmihalyi, 2007), ecological momentary assessment (EMA; Shiffman, Stone, & Hufford, 2008; Stone & Shiffman, 1994), diary methods (including daily diaries; Bolger et al., 2003), ambulatory assessment (Mehl & Conner, 2012), and intensive longitudinal designs (Bolger & Laurenceau, 2013). Some protocols involve daily reports of target behavior for periods as short as a few days or as long as several months. Other protocols ask participants to record their thoughts, feelings, or activities when signaled at random moments throughout the day, or to describe selected events (e.g., social interaction, lies, cigarette smoking) whenever they occur. Yet other methods record activity or physiology continuously.

An obvious reason why these methods have grown in popularity is

technological. Whereas not long ago the questions that everyday experience methods could address were limited, recent developments in miniaturization of digital devices, the accessibility of the Internet and mobile technology, and statistical tools to derive full advantage from the data so provided have engaged the imagination of researchers, allowing them to test hypotheses that earlier researchers could scarcely imagine, much less examine with any precision. Given that the rate of technological advances is, if anything, accelerating, we expect that daily life studies will have an increasing presence in scholarly journals.

There are many different techniques by which researchers can study daily experience. What they share is an appreciation for the complexity, richness, and informativeness of ordinary activity embedded in its natural context. Everyday experience methods are intended to provide detailed descriptions of specific moments, events, or states in a person's life, from which researchers can extract information about “the persistence, cyclicity, change, and temporal structure of thought, emotion, and behavior” (Tennen, Suls, & Affleck, 1991, p. 333), as well as identify situational and dispositional correlates of these patterns. In general terms, we conceive of these methods as a tool for structured contemporaneous self-observation, by which we mean that participants are asked to monitor and describe ongoing activity according to schedules and formats defined and regulated by the investigator. As such, the method is akin to following participants through their day, observing, questioning, or assessing them at relevant points. Everyday experience data sets may include both objective information (e.g., heart rate, recordings of the auditory environment) as well as subjective accounts of experience or other mental processes (e.g., ratings of mood, attentiveness, or sense of self-worth) .

To some extent, the rationale for everyday experience methods is methodological: Descriptions of current feelings and activities minimize, and often eliminate, retrospection bias. Moreover, because data are provided moment by moment, day by day, or event by event, distortions inherent in asking individuals to select, recall, and summarize many events varying in recency and memorability are precluded or at least attenuated (see Schwarz, 2007, 2012 for reviews). Relying on observations that represent behavior across multiple occasions and settings are also likely to enhance the representativeness and generalizability of data.

Everyday experience studies have the further advantage of examining behavior in its natural, spontaneous context. Many years ago, Asch (1952, p. 61)

taught us that perception and action in the social world are determined by the situational context of behavior: “Most social acts have to be understood in their setting, and lose meaning if isolated. No error in thinking about social facts is more serious than the failure to see their place and function.” Although social psychology rightly celebrates itself as the science of situations, we often fail to consider fully the extent to which our phenomena depend on the contexts in which we study them (Cialdini & Paluck, [Chapter 5](#) in this volume; Reis, 2008). McAdams (1995, p. 379) made a similar point about personality, reflecting a Mischelian perspective (e.g., Mischel & Shoda, 1995): “There is no particular reason that the language of nonconditional and decontextualized dispositions should work well to describe constructs that are situated in time, place, and role.” Everyday experience studies permit researchers not only to investigate the operation of social processes in ordinary, self-selected situations but also to characterize those contexts in some detail.

The advantages of the everyday experience approach to social-personality psychology go beyond methodology. As elaborated later in the chapter, these methods allow researchers to develop understandings not easily obtained with other paradigms. For example, everyday experience studies may help establish the real-world prevalence and impact of particular processes and phenomena; may identify situational contexts in which effects are more or less likely to occur; may determine boundary conditions necessary or sufficient for the operation of basic processes; may help distinguish between-person and within-person processes; may identify patterns of cyclicity and covariation among social, cognitive, emotional, and psychophysiological variables; and may clarify their interactions with other, naturally occurring processes. Thus, the procedures we describe in this chapter complement standard research strategies conceptually and methodologically. Everyday experience protocols are founded on the premise that ecological validity matters. We believe that the next generation of social-psychological knowledge will have greater validity, comprehensiveness, and applicability if researchers add everyday experience methods to their toolbox (Brewer & Crano, [Chapter 2](#) in this volume).

This chapter describes everyday experience methods from both conceptual and practical vantage points. We begin with a conceptual rationale, discussing the paradigm's perspective on social behavior and its contribution to social psychological methods. We then review several protocols relevant to research in social and personality psychology and highlight representative studies employing everyday experience methods. The two subsequent sections review practical matters arising in everyday experience research and statistical

techniques for capitalizing on the extensive data sets typically obtained. Finally, we consider the role of everyday experience studies in complementing other methods in programmatic research. Our goal is to foster readers' awareness of the potential benefits of everyday experience methods and to provide a practical guide for incorporating them into a research program.

The Conceptual Rationale for Everyday Experience Methods

Research Aims

Everyday experience studies have three general purposes: establishing the prevalence and/or qualities of phenomena, testing theoretically generated hypotheses and propositions, and serving as a “discovery” technique for generating new hypotheses.

Diary data are commonly used for tallying and describing particular phenomena or constructs. The frequency of given events in everyday life, such as social interaction or exercise, may be estimated from reports of each occurrence, whereas their qualities can be revealed from detailed descriptions contained in each report. Protocols based on sampling units of time, with either random or fixed schedules, yield estimates of the frequency and pattern of given activities in daily life. When recording immediately follows events, retrospection biases are greatly diminished, resulting in relatively accurate accounts, at least from the respondent's point of view.

Here are a few examples. Wheeler and Nezlek (1977) used social interaction diaries, one for each interaction lasting 10 min or longer, to characterize college students' socializing along several dimensions, including frequency, distribution across partners, intimacy, and satisfaction. Carstensen, Pasupathi, Mayr, and Nesselroade (2000) examined how age was associated with positive and negative emotional experiences over a week. Diener, Larsen, Levine, and Emmons (1985) examined the relative frequency of high-and low-intensity emotional states with daily mood reports. Pinkus, Lockwood, Schimmack, and Fournier (2008) recorded instances of upward and downward social comparisons among married couples. Mehl, Vazire, Ramirez-Esparza, Slatcher, and Pennebaker (2007) compared men's and women's talkativeness from ambulatory audio recordings, finding no reliable sex difference in word use. Robinson and Godbey (1997) used daily time logs to describe time allocation among various obligatory and

leisure activities. Adolescent emotional states, and how they vary across settings such as school and home, were studied by Csikszentmihalyi and Larson (1984) with a random time-sampling protocol. Leigh (1993) kept track of alcohol consumption and sexual activity with records completed at the end of every day. Sometimes, researchers combine everyday experience measures with psychophysiological measures, such as by collecting saliva samples in natural environments, examining coregulation of mood and salivary cortisol in married couples (Saxbe & Repetti, 2010), and changes in cortisol associated with separations from a romantic partner (Diamond, Hicks, & Otter-Henderson, 2008).

Beyond their inherent interest value, accurate and detailed descriptive data are essential for theory development. “Before we inquire into origins and functional relations, it is necessary to know the thing we are trying to explain” (Asch, 1952, p. 65). Kelley (1997, p. 166) noted that to develop comprehensive and useful theories, “we need to know more than what are all the possible variations in situations, persons, and interaction. We must also know what the frequent and important variations are.” Similarly, McClelland argued that behavioral frequencies may be the best place for personality theorists to begin (McClelland, 1957). Coming from a background in ethology, Hinde (1995) discussed the importance of descriptive observation in identifying regularities and generating an ordered set of explanatory principles. The biological and physical sciences, after all, began with detailed, systematic descriptions of natural phenomena. For many topics in social-personality psychology, including some for which elegant conceptual models have become standard lore, theoretical refinements may have outpaced basic description. To carve nature at its joints, one must first locate those joints.

Everyday experience data can also be used to test theoretically derived hypotheses. Everyday experience research has been used to establish the functional independence of positive and negative affect – for example, by showing that they are predicted by different types of events (e.g., Clark & Watson, 1988; Gable, Reis, & Elliot, 2000). O’Connor and Rosenblood (1996) used experience sampling to provide evidence for a homeostatic mechanism that maintains social contact at desired levels (i.e., spending more time alone or with others), consistent with individual differences in affiliation motivation. In some cases, time-or event-sampled data are indispensable – for example, for testing hypotheses about cyclicity and variability of emotional states over time (Kuppens, Oravecz, & Tuerlinckx, 2010; Larsen, 1987; Larsen & Kasimatis, 1990), about reactions to stressful events (Stone & Neale, 1984), about stability

in language use across time and social contexts (Mehl & Pennebaker, 2003), about individual differences in self-esteem stability (Kernis et al., 1993), or about cognitions associated with addictive behaviors (Epstein, Marrone, Heishman, Schmittner, & Preston, 2010). In other instances, by evaluating in natural contexts hypotheses also tested with other methods, daily experience studies contribute to the generalizability and validity of a research program (Brewer & Crano, Chapter 2 in this volume).

Hypothesis tests conducted with everyday experience data may serve any of the functions routine in theory-building. For example:

1. *Comparing competing predictions.* Wheeler and Miyake (1992) used a social comparison diary to contrast mood enhancement, which predicts downward comparison following negative mood, with mood consistency, which expects upward comparison. Upward comparison was supported. Fraley, Vicary, Brumbaugh, and Roisman (2011) compared alternative models for change over time in attachment representations. Results supported a prototype model, which assumes that a stable factor underlies any temporary shifts, over a contextual model, which does not assume a stable underlying factor.
2. *Identifying conditions under which processes operate.* Pietromonaco and Feldman-Barrett (1997) showed that the affective consequences of attachment dispositions are most evident during conflict and other anxiety-provoking interactions, as predicted by attachment theory. Downey, Freitas, Michaelis, and Khouri (1998) found that high-rejection-sensitive women reported decreased relationship satisfaction and greater thoughts of ending the relationship than low-rejection-sensitive women, but only following days when they had a conflict with their partner.
3. *Evaluating alternative explanations for a phenomenon.* Bolger and Schilling (1991) compared three explanations for the observed correlation between trait neuroticism and distress. Best supported was the predisposition of persons high in neuroticism to react more strongly to stress. Lucas, Le, and Dyrenforth (2008) compared two explanations for the well-documented association between extraversion and positive affect. Consistent with a temperament model, sociability did not fully account for extraverted persons' greater positive affect.
4. *Unconfounding within-person processes from individual differences.* Reis, Sheldon, Gable, Roscoe, and Ryan (2000) showed that satisfaction

of autonomy, competence, and relatedness needs in everyday activity was associated with greater well-being, over and above the impact of individual differences. Beckmann, Wood, and Minbashian (2010) found that although individual differences in neuroticism and conscientiousness are negatively correlated, neurotic states are positively correlated with conscientious behavior at the within-person level.

5. *Establishing phenomena outside the laboratory context.* As previously shown with laboratory groups of unacquainted individuals, Pemberton, Insko, and Schopler (1996) found that intergroup relations in everyday life are more competitive than are interpersonal relations. Yip (2005) found that ethnic identity became more salient among Chinese Americans in settings that exposed them to the Chinese language, consistent with laboratory findings. Page-Gould, Mendoza-Denton, and Tropp (2008) found that when participants high in implicit prejudice interacted in the lab with someone from another ethnic group, their cortisol reactivity decreased over repeated interactions and they were more likely to initiate intergroup interactions in daily life. In a sample of smokers attempting to quit, Berkman, Falk, and Lieberman (2011) used three weeks' worth of experience sampling data to relate "real-world" cravings and selfregulatory efforts to neural correlates of self-control that had been collected earlier with functional magnetic resonance imagery.
6. *Tracking how behavioral processes unfold over time.* Daily diaries are ideal for charting temporal progressions over time (Bolger et al., 2003). For example, Sbarra (2006) used daily diaries to chart emotional recovery from relationship break-ups. Daily diaries are also useful for studying natural cycles, such as the so-called "day-of-the-week" effect—that positive affect tends to be higher, and negative affect lower, on weekends than weekdays (Stone, Hedges, Neale, & Satin, 1985).

There is no reason why theoretical propositions cannot be evaluated with everyday experience data, much as they are tested in laboratory experiments. Of course, causal inference requires experimental manipulation and random assignment, conditions that can be difficult to achieve in naturalistic contexts (but see Cialdini & Paluck, Chapter 5 in this volume). However, everyday experience methods are meant to complement, not substitute for, experimentation. Establishing a theory's validity, scope, and importance

demands more than simply demonstrating that predictable effects can be evoked under controlled laboratory conditions (Brewer & Crano, [Chapter 2](#) in this volume; Smith, [Chapter 3](#) in this volume). Although the laboratory is best suited for establishing cause and effect, everyday experience studies indicate whether the same effect occurs under voluntary, self-selected conditions, when additional, perhaps unexpected factors come into play. Moreover, the advent of sophisticated procedures for evaluating mediational models and propositions (see Bolger & Laurenceau, 2013; Judd, Yzerbyt, & Muller, [Chapter 25](#) in this volume) makes everyday experience research increasingly useful for theory building.

Finally, everyday experience methods sometimes play a serendipitous role in research. McGuire ([1997](#)) asserted that the task of generating creative hypotheses is perhaps the most difficult and poorly understood step in the research process. Among various strategies he recommended are several techniques for sensitively observing natural occurrences. By assembling large data sets with multiple predictor, outcome, and moderator variables, diary studies lend themselves to exploration – “scouting out” new hypotheses (Mortensen & Cialdini, [2010](#)) – and the prospect of theory-advancing insights. Hypotheses so generated can be tested in subsequent studies. For example, Clark and Watson ([1988](#)) found that whereas physical illness had been linked only to negative affect in between-subjects analyses, within-person analyses linked illness to both high negative and low positive mood, an effect replicated in later studies. Or consider a study by Mohr, Armeli, Tennen, Carney, Affleck, & Hromi ([2001](#)), in which it was observed that alcohol consumption was likely to occur in the presence of others who are also drinking. A laboratory study might not have differentiated social and solitary drinking, lessening its usefulness for understanding how people actually drink.

In sum, everyday experience methods are adept at both description and hypothesis testing. Their main advantage is enhanced ecological validity, an important criterion that too often receives short shrift in social and personality psychology (but see Brewer & Crano, [Chapter 2](#) in this volume). Because phenomena are assessed within natural contexts, artifacts attributable to setting or other incidental aspects of the research process are greatly reduced. Although maximum internal validity is sacrificed, the emphasis on contemporaneous reports repeated over time and context minimizes retrospection and other forms of self-report bias, making the loss of internal validity less than with other nonexperimental methods.

Conceptualizing Everyday Experience

Domains of experience. Although researchers often choose their tools for methodological and practical reasons, paradigms also reflect conceptual distinctions. Reis (1994) described three domains of inquiry, each scrutinizing a phenomenon or process from a distinct perspective. Research programs that incorporate all three perspectives are likely to be most informative.

One perspective, termed *exemplary experience*, consists of studies in which behavior is observed in specific, restricted, or otherwise special settings. This includes laboratory experiments, in which behavior is observed under controlled conditions, and observational studies, which are carried out in uniform, often intrinsically relevant settings such as playgrounds, worksites, and kitchens (Cialdini & Paluck, Chapter 5 in this volume). One meaning of the word “exemplary” is “commendable; worthy of being imitated.” Research participants aware of being monitored may exhibit optimal rather than typical performance (Ickes & Tooke, 1988). Well-known processes such as impression management, social desirability, demand characteristics, politeness, evaluation apprehension, and the desire to be helpful or agreeable may impel behavior that would differ away from the scrutiny of experimenters. Thus, in the lab, a bully might react calmly to a confederate's provocation, knowing that his responses were being recorded.

Particular settings may induce optimal behavior even when participants are unaware or unconcerned about being observed. Many contexts, notably including research laboratories, elicit polite, formal, cooperative, or thoughtful behavior that departs from behavior displayed in everyday settings. Marital interaction observed in the laboratory may differ from marital interaction at home (Larson, Richards, & Perry-Jenkins, 1994); for example, in the laboratory, participants rarely leave the room when asked to self-disclose or carry out an unpleasant task, as spouses sometimes do at home. Situational cues provided by research laboratories have not been studied extensively (but see Shulman & Berman, 1975), although it seems clear that the laboratory setting itself may prime certain expectations and scripts (e.g., scientific legitimacy, serious purpose, suspicion about possible deception, concerns about being observed, the need for attentiveness), all of which may affect participants' thoughts and behavior in both intentional and unintentional ways.

A rather different definition of exemplary is “serving as an illustration; typical.” This definition refers to the assumption, inherent in laboratory work,

that observed behavior typifies participants' natural responses to the conditions being studied. However, the cardinal rule of experimentation – carefully controlled context – necessarily constrains the range of possible behavior (Smith, [Chapter 3](#) in this volume). Social judgment studies, for example, rarely allow participants to change the topic, qualify their answers, or simply say nothing, as people so often do in real life. Seasoned experimenters know that slight changes in context may beget large changes in behavior. To be sure, careful control of context gives experimentation its exceptional power for validating causal hypotheses. However, the contextual background customarily receives far less attention than does the manipulation and outcome, even though that background may embody influential boundary conditions (i.e., moderators).

In short, although there are good reasons for studying exemplary settings, knowledge garnered from such investigations may be limited. A controlled setting adds precision to what can be observed but “inevitably fails to incorporate the broader pattern of behaviors and contexts that make up daily life” (Funder, [1991](#), p. 36). Such findings are therefore likely to be enhanced by complementary insights from other approaches.

The second perspective, also familiar to social-personality psychologists, is *reconstructed experience*. Here, participants are asked to evaluate, summarize, or otherwise describe in questionnaire or interview format their experiences with specific entities or in particular situations. Self-generated global assessments, a valuable source of data about perceptions, often differ from online experience (Kahneman, Fredrickson, Schreiber, & Redelmeier, [1993](#)) because of the many processes that influence encoding, storage, retrieval, and evaluation of episodic memories (discussed later in the chapter). Beyond limits in people's ability to recall and summarize past experiences, even spanning brief intervals, motivated processes such as cognitive efficiency and self-esteem maintenance commonly transform event-by-event or moment-by-moment memories (Schwarz, [2007](#), 2012). Responses to global questions are therefore better considered as reconstructed interpretations of personal experience than as direct accounts of that experience.

Comparison of contemporaneous and reconstructed evaluations of the same experience may illuminate these processes. In some cases, as in Stone, Schwartz et al.'s (1998) comparison of momentary and every-other-day retrospections about coping, there is little or no correspondence.¹ In other cases, divergences highlight processes of interest. For example, Redelmeier and Kahneman ([1996](#)) showed that recall of pain from unpleasant medical procedures was based

primarily on the most intense level of pain experienced and the most recent, or end, level (termed the “peak-end” rule). Updegraff, Gable, and Taylor (2004) found that people dispositionally high on approach motivation were more likely than those low on approach motivation to base their judgments of daily satisfaction on momentary positive emotions experienced throughout the day. Reconstructed impressions also may reflect implicit theories about events rather than their actual content (Ross, 1989), perhaps to bolster current beliefs. For example, in a four-year longitudinal study, Sprecher (1999) found that although yearly ratings of love in successful relationships were stable, spouses supported the belief that love had grown over time by lowering recollections of prior love. Similarly, men and women differ in retrospective but not in momentary reports of emotional experiences (Barrett, Robin, Pietromonaco, & Eysell, 1998), suggesting that sex differences in emotion may be more a matter of stereotype-guided recollection than of actual experience.

The influence of episodic or momentary experiences on general impressions represents an important substantive question for social-cognition research (Carlston & Smith, 1996). How, for example, do fluctuations in affect across repeated interactions with a partner evolve into global evaluations (Campbell, Simpson, Boldry, & Rubin, 2010)? How do momentary moods contribute to global feelings of life satisfaction or dissatisfaction (Diener, 1996; Updegraff et al., 2004)? To understand reconstructions as a product of cognitive and self-serving transformations enacted on actual experience, it is useful to investigate the transformational process directly, by comparing global impressions with data from the third domain, that of ongoing experience.

The third domain, *ongoing experience*, focuses on direct, usually immediate reports of everyday experience. Verifying causal antecedents with maximum internal validity is generally not the overriding concern; rather, these studies aim to maximize external validity by examining specific processes or phenomena within the stream of routine, voluntary activity. With suitable analysis, they permit specification of contexts in which target behaviors do and do not occur, they identify natural patterns of variation in target behaviors and covariation with predictors and spontaneous consequences, and they document the prevalence and nature of phenomena.

What type of research fits this category? Generally, these studies share an interest in the ongoing, often mundane moments and occurrences of everyday life. They concern the diverse feelings, thoughts, and activities that occur spontaneously, filling people's waking time and occupying most of their

conscious thoughts and attention. Daily life events have a structure and rhythm of their own, sometimes variable and fleeting, at other times stable and continuous. Some daily events are vivid and arousing, others are mundane and inconsequential. The central assumption behind this approach is that all such experiences matter and, when examined carefully, may provide valuable insights about human behavior.

Because of its unique perspective, studies of ongoing experience have contributed insights unattainable with traditional methods in many domains of social-personality psychology. For example, many studies have examined patterns and correlates of day-to-day and within-day variations in mood (e.g., Clark & Watson, 1988; David, Green, Martin, & Suls, 1997; Kuppens et al., 2010). Other studies have looked at cognitive activity and motivation among high school students (e.g., Moneta & Csikszentmihalyi, 1996) by randomly timed momentary reports. DePaulo, Kashy, Kirkendol, Wyer, and Epstein (1996) had participants keep daily logs of their lies, whereas Wheeler and Nezlek's (1977) participants completed records of their social interactions. Fleeson (2001) examined within-person variability in states and behavior relevant to the Big-Five personality traits. Other research has explored daily experiences of racial discrimination (Ong, Fuller-Rowell, & Burrow, 2009; Swim, Cohen, Hyers, Fitzgerald, & Bylsma, 2003), aggression (Pond, DeWall, Lambert, Deckman, Bonser, & Fincham, 2012), and intimate partner violence (Finkel et al., 2012). Momentary or daily reports are commonly used to identify determinants and health consequences of coping and stress (e.g., Bolger, DeLongis, Kessler, & Schilling, 1989; Repetti, 1989) as well as social support (e.g., Gleason, Iida, Shrout, & Bolger, 2008). Researchers have also studied fluctuations in self-concept and self-focused attention (Campbell, Chew, & Scratchley, 1991; Hormuth, 1990; Kernis, Cornell, Sun, Berry, & Harlow, 1993) and self-esteem (Knee, Canavello, Bush, & Cook, 2008). Finally, everyday experience studies are increasingly used to test meditational hypotheses. For example, in a 42-day daily diary study, Laurenceau, Barrett, and Rovine (2005) found that perceived partner responsiveness mediated the effects of selfdisclosure on intimacy. A more diverse set of topical reviews can be found in Mehl and Conner (2012).

Our premise is that each domain offers a different but no less valuable perspective on a given process or phenomenon. Studies of exemplary experience are informative about behavior in particular, well-specified contexts, and they help establish the causal impact of those contexts. Studies of reconstructed experienced tell us how people understand their lives and activities. Studies of everyday experience provide insights about thoughts, feelings, and activities that

occur in natural settings. Irrespective of methodological considerations, important conceptual benefits accrue from including all three domains in a research program.

Major vs. minor events. Researchers have had long-standing interest in the impact of major life events, such as marriage, divorce, bereavement, the birth of a child, major illness, and employment changes, on emotional well-being, health, and social activity (e.g., Kessler, 1997; Lucas, 2007). Everyday experience research adopts a different perspective, exploring the impact of minor, or mundane, events on the same general outcomes, based on the assumption that ordinary evenings spent in quiet conversation with a partner or irritating days at work may also be influential, especially when recurrent. Enduring patterns of emotion or interaction may matter in the long run, even if any single episode is negligible – habitual types of communication between spouses, ongoing patterns of interaction among teachers and students, or coworkers, supervisors, and employees, or chronic styles of selfregulation, for example.

That daily hassles influence health and well-being has been shown in several studies (e.g., DeLongis, Folkman, & Lazarus, 1988; Stone, Neale, & Shiffman, 1993). Mundane daily events, such as helping a friend move or riding on a crowded bus, relate to daily mood and symptoms, even after controlling for major life events (Clark & Watson, 1988). Although the effects of major events are undeniable, their significance may be limited by infrequency and because their impact tends to subside over time, as research on hedonic adaptation to life events has shown (e.g., Fujita & Diener, 2005; Wortman & Silver, 1989). For example, Suh, Diener, and Fujita (1996) found that only life events during the past three months mattered for emotional well-being. In contrast, mundane events occur far more often: The vast majority of affects experienced in everyday life are low intensity (Diener et al, 1985), and the vast majority of interactions are routine and superficial, even in very intimate relationships (Hays, 1989). Their impact is therefore more likely to be evident in patterns over repeated instances.

The study of minor life events and states lends itself well to the everyday experience approach. By definition mundane, these events tend to be unmemorable, compromising retrospective methods. Moreover, people tend to have difficulty perceiving regularities or cyclical variations within a series of relatively inaccessible events, so that global self-reports are likely to be uninformative or misleading. In contrast, repeated, contemporaneous reports of even the most forgettable feelings or events allow researchers to identify

whatever meaningful patterns may exist and whatever consequences they may have.

Everyday experience studies may also reveal repercussions of major events and chronic stressors. Certain life events, such as divorce, spousal loss, or a heart attack, may have their greatest impact by disrupting everyday routines and altering ongoing mood and thought (Caspi, Bolger, & Eckenrode, 1987). Similarly, several studies have found that daily stressors account for differences in daily well-being associated with more chronic stressors, such as low socioeconomic status (Gallo, Bogard, Vranceanu, & Matthews, 2005) and chronic racial discrimination (Ong et al., 2009). Retrospective impressions of major life events may elicit naive theories about the presumptive impact of transformative events rather than objective accounts (Ross, 1989). On the other hand, their actual impact can be established with diary records, in longitudinal (i.e., comparing pre-event and post-event experiences) or cross-sectional designs.

Distinguishing between-person and within-person effects. Whereas most researchers are aware of the need to differentiate between-person and within-person effects for statistical reasons, the conceptual distinction is often overlooked. Consider the following hypothetical studies of trust and selfdisclosure. In Study 1, 100 participants are asked how much they trust their best friend. Selfdisclosure is also rated, yielding a correlation of .50. Participants in Study 2 are asked to rate trust and selfdisclosure for each of 10 different friends and acquaintances. Correlations between trust and selfdisclosure are computed for each participant across their 10 ratings and then averaged across participants. The resulting correlation is also .50. Do these two correlations document the same phenomenon?

No. The first correlation indicates that people high in trust (particularly with best friends) also tend to be high in selfdisclosure. Several explanations are possible, prominent among them the dispositional alternative, which is that certain traits predispose people to be trusting and also to be open with others. The second correlation shows that the more one trusts a particular other, the more selfdisclosing one is with that person. Because the correlations are computed within-person, dispositional explanations are irrelevant, supporting an interpretation that trust and selfdisclosure covary as a function of relationship qualities.

Conceptually, it is easy to see why these levels of explanation are independent. Correlations computed between persons ask whether persons

scoring high on one variable tend to score similarly high on another variable. This design is appropriate when the research question involves dispositional processes, but when covariation across conditions, circumstances, or occasions is of interest, within-person correlations are more suitable. Methods for studying within-person processes are less well known than between-person methods are, so they are sometimes investigated (erroneously) in between-person designs (Gable & Reis, [1999](#)).

Two examples may highlight implications of this distinction. Epstein ([1983](#)) found a positive correlation between daily reports of sadness and anger with between-person data, suggesting that people predisposed to experience one emotion were also more likely to experience the other emotion. On the other hand, the average within-person correlation for the same variables was negative, indicating that individuals were less likely to feel anger on days they felt sad. (Of course, both findings are readily integrated theoretically.) Second, in between-person analyses, Emmons ([1991](#)) found no significant difference in the extent to which personal strivings moderated reactivity to different daily events, but within-person analyses indicated that achievement-oriented participants were more reactive to good achievement events, whereas affiliation-oriented participants were more affected by interpersonal events.

Within-person processes are particularly prominent in two research areas: relationships and personality. In the former, theories often address how behavior varies as a function of the relationship context (Reis, Collins, & Berscheid, 2000) – in other words, how behavior varies when a person is interacting with one or another partner, depending on the nature of their relationship – a type of conceptual question well-suited to daily experience methods. As for the latter, personality theorists increasingly recognize that personality reflects “not just a person's average way of behaving but also a person's range of behavior, including the amount of variability” (Fleenon & Nofle, [2012](#), p. 528; see also Fleenon, [2001](#)) and the nature of the situations that foster such variability (Mischel & Shoda, [1995](#)). Daily experience methods are ideal for elucidating how traits affect variability of behavior across situations, known as Person x Situation interactions (Reis & Holmes, [2012](#)) .

The Methodological Rationale for Everyday Experience Methods

Most researchers agree that cognitive and motivational processes tend to bias

responses to survey questions; as Stone and Turkhan (2000, p. ix) put it, “It is naive to accept all self-reports as veridical.” Consider questions like, “All things considered, how satisfying are your interactions with friends?” or “How stressed have you felt during the past week?” Respondents must retrieve from memory the full set of qualifying events, selecting from among them the most relevant subset. Features germane to satisfaction or stress must then be remembered, rated, and combined with some sort of decision rule into an overall impression. Of course, no one carries out this exercise completely. Instead, heuristics and other cognitive shortcuts allow people to respond quickly and efficiently, albeit with varying accuracy (Fiske & Taylor, 2007; Kahneman, Slovic, & Tversky, 1982). Adding these to the inevitable decay of memory leads to the unsurprising conclusion that recollection is often flawed.

Wentland (1993) meta-analyzed studies that used objective criteria to evaluate the accuracy of self-reports. (Attitudes, feelings, and impressions were excluded because they cannot be verified independently.) Across a diverse list of both sensitive and nonthreatening behaviors, accuracy ranged from 23% to 100%. The main factor influencing accuracy was information accessibility, operationally defined with specific variables such as length of recall period, event salience, and question and response clarity. Wentland's analysis, which is consistent with other analyses of self-report bias (e.g., Schwarz & Sudman, 1996), suggested that concrete, specific questions close in time to the event tend to be most accurate. The accuracy of subjective variables likely follows the same general principle.

Although much inaccuracy is random, systematic errors have dominated the field's attention (Schwarz, Groves, & Schuman, 1998), with reconstruction often described as a heuristic-driven process. Among the more important heuristics are the following:

- *Recency*. The more recent the event, the better it is recalled and the more likely it is to influence retrospection. For example, Bernard, Killworth, Kronenfeld, and Sailer (1984) reported a series of experiments examining informant accuracy in recalling social network or communication contacts. Overall, fewer than half were remembered correctly, but more recent events were recalled better than less recent events were. Even memory for relatively distinctive events degrades daily, for up to two months, between the event and its recall (Skowronski, Betz, Thompson, & Shannon, 1991). Also, end-of-day summaries of mood are sensitive to recent occurrences (e.g., Stone et al., 1993), consistent with the “peak-end” rule (Kahneman et

- al., 1993).
- *Salience*. More distinctive events, in terms of intensity, emotionality, unusualness, or personal significance, tend to be more influential. For example, summary ratings of daily or weekly emotion reflect moments of peak intensity more than they reflect average levels (Parkinson, Briner, Reynolds, & Totterdell, 1995; Thomas & Diener, 1990). Individuals also retrospectively recall greater levels of pain if their daily experiences of pain were more variable (Stone, Schwartz, Broderick, & Shiffman, 2005). Similarly, general impressions of relationships reflect salient emotional interactions more than less-emotional interactions (Pietromonaco & Feldman-Barrett, 1997).
 - *Sense-making*. Events tend to be interpreted in light of later developments or to conform to implicit theories and beliefs (Ross, 1989). For example, women's recollections of their menstrual symptoms during a four-to six-week span better resembled their general beliefs about such symptoms than their daily symptom reports (McFarland, Ross, & DeCourville, 1989). People high in neuroticism recall more negative emotion than their daily diaries indicate, whereas people high in extraversion tend to overreport positive emotion (Feldman-Barrett & Pietromonaco, 1997).
 - *State of mind*. Retrospections, especially of affect and attitude, may reflect mood at the time of report (Blaney, 1986). For example, global summaries of mood over various intervals tend to resemble current mood (Parkinson et al., 1995; Stone et al., 1993), perhaps through availability biases. Similarly, certain motives and perceptions are assessed more accurately when relevant states of mind are activated (McClelland, Koestner, & Weinberger, 1989), for example, ratings of perceived social support following emotional arousal (Clark, Fitness, & Brissette, 2001).

These and related biases (Stone, Shiffman, Atienza, & Nebeling, 2007; see Schwarz, 2012 for a review) provide the prime methodological rationale for event-sampling methods that require instantaneous, in-the-moment reports (sometimes called real-time data capture; Stone et al., 2007). Note that all of the sampling schemes discussed later in the chapter are amenable to instantaneous reports. Researchers might keep in mind that although momentary reports are by definition more accurate, they are not necessarily more important than recollections, which may relate better to certain outcomes. Comparisons of event-sampled and retrospective data may be helpful in keeping this distinction clear .

Diary Data as Self-Reports

Some researchers use experience sampling to avoid the shortcomings of self-report, noted earlier. Others, however, criticize those experience-sampling studies that ask participants to report on their experiences because their answers are still self-reports. Although literally true, this criticism overlooks the key difference between global recollections and momentary, usually contemporaneous accounts: Everyday experience methods obviate or at least minimize many (though not all) specific biases that plague global self-reports.

Several sources support this premise. Survey research has found that decompositional approaches, which ask circumscribed questions about the smallest possible units, are superior to open-ended, global questions, which invite heuristic processing (e.g., Menon, 1997). Similarly, observer ratings tend to be more accurate the less global and the more concrete they are and the more they focus on discrete behaviors rather than general impressions (Heyman, Lorber, Eddy, & West, Chapter 14 in this volume; Ritter & Langlois, 1988), a principle that likely generalizes to self-reports. In this regard, Penner, Shiffman, Paty, and Fritzsche (1994) showed that within-person estimates of mood variability across diary reports were free of response-bias artifacts. The advantages of a componential approach dovetail with the benefits of aggregation across time or situations (e.g., computation of composites or covariation analysis). Aggregated data generally identify stable patterns of behavior better than do data from single situations or assessments, which may be influenced by atypical or random factors (Epstein, 1983).

Verifying the advantages of event-sampling approaches is complicated by the difficulty of defining appropriate criteria. Criteria that are themselves global representations (e.g., subjective well-being or school achievement) may relate better to other global variables as a result of shared method variance. The relative accuracy of different methods can be assessed by comparing independent observations. For example, Conrath, Higgins, and McClean (1983) asked participants to keep track of 100 interactions with coworkers. Concurrence between two individuals that an interaction had taken place was nearly twice as great with diaries than with a global “communications” questionnaire. Reis and Wheeler (1991) reported several studies showing high correspondence between roommates in reporting that an interaction had taken place. As for more subjective variables, earlier we discussed evidence that event-sampled data are substantially less influenced by self-report biases than are global self-reports. Independent verification of subjective variables is more problematic, because

disagreement may denote differences of opinion or perspective rather than inaccuracy. Reis, Senchak, and Solomon (1985) found, however, that participants' ratings of intimacy for a single interaction were highly correlated with judgments by independent observers.

To be sure, event-sampled data are self-reports, and as such are not distortion free. Nevertheless, it seems clear that event-sampling protocols characterize ongoing experience with substantially greater accuracy than global self-reports do. Self-reports are unlikely to disappear anytime soon from social-psychological research, if only because many important phenomena – affects, attitudes, cognitions – intrinsically depend on the individual's interpretation of his or her circumstances. As Kagan (1984, p. 241) noted, “The child's personal interpretation of experience, not the event recorded by camera or observer, is the essential basis for the formation of and change in...beliefs, wishes, and actions.” Everyday experience methods offer a substantially improved way of obtaining self-reports.

Types of Everyday Experience Protocols

In this section we describe three protocols for everyday experience studies, a typology proposed by Wheeler and Reis (1991). The three models differ primarily in the sampling frame used to obtain data. These distinctions are not merely procedural details; each protocol is tailored to fit particular operational circumstances and theoretical goals, and findings may depend on the choice of method.

Interval-Contingent Recording

With this method, sometimes called *time-contingent recording*, participants report their experiences at regular, predetermined intervals. Typically, these intervals represent theoretically or logically meaningful units of time, such as the end of each day or every four (waking) hours. Interval-contingent reports are commonly used in two ways: to describe events that have transpired since the previous report or to depict circumstances at the moment of recording. Either way, it is important to space intervals reasonably. If the range is too great, natural cycles (e.g., diurnal rhythms) may be obscured, or important intervening events may be excluded. If the span is too short, the signal-to-noise ratio may be small and the burden on participants may be excessive, potentially reducing compliance and attentiveness.

By far the most common interval sampling unit is the day. A good example is the National Study of Daily Experiences (Almeida, 1997), in which a national sample of 1,484 adults completed telephone interviews about their daily experiences on eight consecutive evenings. The Day Reconstruction Method (Kahneman, Krueger, Schkade, Schwarz, & Stone, 2004) asks participants to systematically reconstruct the prior day's activities and experiences, using procedures typical of time-budgeting and experience-sampling studies to minimize recall bias. Other examples include so-called daily diaries – once-daily reports, usually for periods ranging from one week to as long as several months, examining, for example, mood and events (Clark & Watson, 1988), conflicts, coping, and distress (Bolger & Zuckerman, 1995), lies (DePaulo et al., 1996); time allocation across activities (Robinson & Godbey, 1997), health maintenance practices and symptoms (Lawrence & Schank, 1995), intergroup contact (Page-Gould, Mendoza-Denton, & Tropp, 2008), racial discrimination and stress (Ong, Fuller-Rowell, & Burrow, 2009), aggression and intimate partner violence (Finkel, DeWall, Slotter, McNulty, Pond, & Atkins, 2012), motives (Woike, 1995), social support (Gleason, Iida, Shrout, & Bolger, 2008), sex (Birnbaum, Reis, Mikulincer, Gillath, & Orpaz, 2006), or self-focused attention and coping (Wood, Saltzberg, Neale, Stone, & Rachmiel, 1990). The value of once-daily recording is consistent with intuition and empirical evidence that sleep-and-awakening provides a discrete break in biological and psychological cycles (Williams, Suls, Alliger, Learner, & Wan, 1991).

Other fixed intervals used in social-personality research include weekly assessments of attachment representations over one year (Fraley, Vicary, Brumbaugh, & Roisman, 2011), aggression measured three times each week for 25 days (Pond et al., 2012), self-esteem ratings twice a day, at 10 AM and 10 PM (Kernis et al., 1993), mood ratings at 9 AM, 1 PM, 4 PM, and 7 PM (Hedges, Jansdorf, & Stone, 1985), physical symptoms described at noon, dinnertime, and bedtime (Larsen & Kasimatis, 1991), psychological states associated with the Big 5 assessed every 3 hours from noon to midnight (Fleeson, 2001), and 50-second speech samples collected every 9 minutes (Mehl & Robbins, 2012). Although intervals are often chosen intuitively or for convenience, spectral analysis can identify repetitive cycles in behavior or states, which would allow researchers to select optimally spaced intervals (Larsen, 1990). For example, a weekly mood cycle is well documented, peaking on weekends (Larsen & Kasimatis, 1990).

Signal-Contingent Recording

With signal-contingent reports, participants describe their activity at the moment when a signal is delivered, most often by smartphones, pagers, or other programmable devices. Signals may follow fixed or random schedules, or a combination of the two (e.g., random signals within preset blocks, such as every two hours). If schedules are random (or at least unrelated to participants' activities) and if participants comply with instructions to respond immediately, signal-contingent data can be used to estimate the prevalence and distribution of activities and states over time. Signals that are regular, predictable, or subject to the participant's choice may yield unrepresentative data because of self-selection or regularities in activity or states of mind. Randomness is also desirable so that participants cannot modify their activities in anticipation of a signal.

The original and best-known example of signal-contingent recording is the experience sampling method (ESM; Hektner et al., 2007), developed by Csikszentmihalyi and colleagues and first used to describe the “ecology of adolescent experience” (Csikszentmihalyi, Larson, & Prescott, 1977). ESM reports of self-described activity, thoughts, and feelings are cued by an electronic signal programmed to occur at random (and hence unpredictable) moments throughout the day. In a typical study, Delespaul (1995) divided the time between 7:30 AM and 10:30 PM into ten 90-minute intervals, randomly beeping participants once during each period, subject to the constraint that beeps be at least 15 minutes apart. Questions may be rating scales or open-ended. Although early ESM studies relied on pagers and paper booklets, reflecting the technology available at the time, more recent work typically uses digital presentation and recording (e.g., smartphones or PDAs), which has the added benefit of verifying that data entry coincided with the signal. (Verification of response time is discussed later.)

A slightly more general version of ESM is ecological momentary assessment (EMA), developed by Stone and Shiffman (1994). EMA has been used to link ESM-like self-reports with ambulatory measurements of physiological states in natural environments. For example, Gallo *et al.* (2005) related mood ratings to ambulatory blood pressure readings taken at random intervals over two days, whereas Lane, Zareba, Reis, Peterson, and Moss (2011) related mood ratings to ventricular repolarization assessed from continuous Holter monitor readings. Rapid technological advances in portable physiological monitors and in the accessibility and convenience of smartphones and other types of handheld digital devices for administering complex protocols have allowed the ESM and EMA to

become increasingly common (Intille, 2012; Miller, 2012). Mehl and Conner (2012) provide a comprehensive review of these tools and their potential uses.

The flexibility of signal-contingent recording for tracking diverse, naturally fluctuating phenomena has made it especially popular for studies of mood and health-related phenomena, both because they tend to vary in theoretically interesting ways in everyday circumstances and because recollections tend to deteriorate rapidly (Conner & Barrett, 2012; Stone & Shiffman, 1994). Other topics studied by signal-contingent recording include daily experiences of desire and self-control (Berkman et al., 2011; Hofmann, Baumeister, Förster, and Vohs, 2012), adolescents' thoughts and feelings (Csikszentmihalyi & Larson, 1987), coping with everyday stressors (Schwartz, Neale, Marco, Shiffman, & Stone, 1999), social rejection (Eisenberger, Gable, & Lieberman, 2007), self-relevant cognition (Hormuth, 1986), personal strivings (Emmons & King, 1988; 1989), approach-avoidance motivation (Updegraff et al., 2004), mental life and activity among people with schizophrenia (Delespaul, 1995), adolescent-parent interaction (Larson & Richards, 1994), and personality processes (Brown & Moskowitz, 1997; David et al., 1997).

Event-Contingent Recording

Event-contingent recording requires a report whenever events occur matching a predetermined definition. The Rochester Interaction Record (RIR), developed by Wheeler and Nezlek (1977), calls for completion of various rating scales and descriptive items after every social interaction lasting at least 10 minutes. For instance, a longitudinal study by Reis, Lin, Bennett, and Nezlek (1993) examined change and consistency in social interaction from the college years to adult life approximately 10 years later. Using the RIR to assess social interaction at both times, they found that over time, socializing became more focused on opposite-sex partners and generally more intimate but no more satisfying. Similar protocols have been used to study conversations (Duck, Rutt, Hurst, & Strejc, 1991), lies (DePaulo et al., 1996), adolescent conflict (Jensen-Campbell & Graziano, 2000), agentic and communal dimensions of social interaction (Moskowitz, 1994), relationship-contingent self-esteem (Knee, Canavello, Bush, & Cook, 2008), social comparison (Pinkus et al., 2008; Wheeler & Miyake, 1992), self-presentation (Leary, Nezlek, Downs, Radford-Davenport, Martin, & McMullen, 1994), smoking (Shiffman, Paty, Gnys, Kassel, & Hickcox, 1996), social and solitary drinking (Mohr et al., 2001), food and drink consumption (Decastro & Pearcey, 1995), and sexual activity (Birnbaum et al., 2006).

Key to event-contingent recording is unambiguous definition of the events to be described, as well as timeliness. If participants are permitted to select events, or if it is unclear whether a given event should be reported, systematic distortions are possible. Event-contingent recording is useful when events are relatively low-frequency (in which case signal-contingent methods would capture few instances), especially when subtypes are of interest (e.g., same-sex vs. opposite-sex interaction; comparisons of drinking alone versus with others). Event criteria often take participant burden into account, such as by requiring that only certain subtypes be recorded (such as interactions involving specific others or lasting 20 min or longer). Moskowitz and Sadikaj (2012) provide a more detailed discussion of event-contingent recording.

Comparison of Protocols

Selection among these three protocols reflects several considerations, including research goals, the relative frequency with which the central processes occur or vary, the time frame in which report accuracy is likely to degrade, and participant burden. Interval-contingent methods are usually chosen to examine fluctuations over time (e.g., day-by-day variations in mood, perceived competence, or alcohol consumption). Because the time span from one record to the next is constant, interval-contingent methods are ideally suited to time-series analysis, for which the irregular gaps of signal-contingent and event-contingent recording add complexity (discussed later in this chapter). Interval sampling is also appropriate when the time unit has inherent meaning. For example, experience is often summarized in single-day units (e.g., “How was your day?”), or one might examine hourly variations in cognition and affect among psychotherapists.

Interval-contingent recording has the further advantage of minimizing participant burden, inasmuch as intervals are relatively lengthy and predictable, facilitating data collection over longer spans. Because respondents have some control over the exact moment of data entry, intrusiveness is substantially less than with protocols requiring immediate reports whenever a signal or relevant event occurs – be it during a meeting, party, or while preparing a soufflé. Participants need simply to remember to complete a record at the appropriate time, or be available for regular, prearranged telephone interviews (e.g., Almeida, 1997).

A major disadvantage of interval-contingent recording is that events may be removed in time from reports, adding possible retrospection bias. Furthermore,

target variables may have fluctuated or occurred more than once during an interval, making them difficult to describe. Thus, the cognitive biases that daily experience studies are designed to minimize may affect interval-contingent responses, depending on the time between event and description and on memorability. Weekly reports of social activity probably would contain substantial inaccuracies, for example, whereas hourly time budgets and weekly exercise reports would probably be reasonably precise (Gunthert & Wenze, 2012). Of course, retrospection bias is less of a concern for variables that inherently involve impressions aggregated over time and experience (e.g., “how stressful was work today?”). Repeatedly assessing behavior or feelings at fixed times or in the same setting may introduce systematic bias. For example, people may be lethargic at bedtime or may exclude certain information when their spouse is present.

Signal-contingent recording tends to be more intrusive than event-contingent recording. Signals, even when inaudible to others (e.g., using vibration alerts), may interrupt ongoing activity, whereas event-contingent reports are provided after an event's natural completion. For practical reasons, most signal-contingent protocols allow participants to postpone responding at inopportune moments or to switch off the signaling device when they are unwilling to be disturbed. This concession works against the necessary randomization and representativeness, however. On the other hand, because some events vary within themselves – interactions that begin superficially and become progressively more intimate, for example – event-contingent recording may involve some degree of aggregation within a single report.

Either signal-or event-contingent formats are preferable for determining the prevalence of events and states in everyday life. However, event-sampling is effective only when the event can be defined unambiguously, so that ambiguous instances are not overlooked. Signal-contingent methods are preferable for assessing the relative frequency of activities (e.g., time budgets; Robinson & Godbey, 1997) or states (e.g., intense vs. mild emotions; Diener et al., 1985). Otherwise, participants would need to complete a record for almost everything they did – an onerous task. Signal-and interval-contingent data collected electronically have the added advantage of recording the time of signal and response, which can verify compliance with the sampling schedule. Whether participants have faithfully followed an event-contingent scheme can be ascertained only from informants or independent observers.

Finally, event-contingent sampling is most effective when the events are rare,

or when variations within a class of events, some of which are rare, are of interest. For example, even with 7–9 signals per day for 1 week, high school students reported an average of only 4.16 instances of studying (Wong & Csikszentmihalyi, 1991), a small number for accurately estimating mental states while studying. Similarly, event-sampling with college students yielded an average of only 3.90 interactions with romantic partners per week (Tidwell, Reis, & Shaver, 1996). Random signals would be unlikely to obtain more than one or two instances, unless data collection continued for a prohibitively long time. Cross-matched categories – studying with friends as opposed to alone – are even rarer. In that comparisons of mental states during different activities (e.g., studying vs. socializing or same-sex vs. opposite-sex interaction) are likely to interest researchers, methods that sample rare events efficiently are often needed. Table 15.1 summarizes this discussion, listing advantages and disadvantages of each protocol.

Table 15.1. Comparison of Everyday Experience Protocols

Protocol	Situations Favoring Protocols
Interval-contingent	<p>When susceptibility to retrospection bias is low</p> <p>To minimize participant burden</p> <p>When aggregated time intervals are inherently meaningful</p> <p>To conduct time-series analyses and evaluate cyclical patterns of variation and covariation</p>
Signal-contingent	<p>When susceptibility to retrospection bias is high</p> <p>To establish the relative frequency and distribution of different activities or mental states</p> <p>To compare different domains of activity or mental states during different activities</p> <p>To verify time of recording</p>

TO VERIFY TIME OF RECORDING

Event-contingent

When susceptibility to retrospection bias is high

When interested in a specific class of events or states, especially rare, clearly defined events

To compare infrequent variations within a class of events

When accounts of many episodes of the event or state under investigation are needed

Pragmatic Considerations in Everyday Experience Research

As with all methods, event sampling involves many mundane decisions and practices that, if mishandled, can adversely affect research outcomes. Often these decisions reflect tradition (“that’s how we’ve always done it”), expedience, or human-participant realities, although it is to be hoped that these trade-offs do not compromise conceptual priorities. First-time forays into everyday experience methods can be daunting, because studies tend to be labor-intensive for participants and researchers alike. Although it is difficult to specify universal principles and procedures, because every study has different theoretical aims and pragmatic issues, we next offer general guidelines that readers may find helpful in designing and planning research .

Designing the Protocol

The choice among interval-, signal-, and event-contingent schemes should be based on two considerations: the type of question(s) being studied and the incidence of target behaviors. Earlier we discussed appropriate and inappropriate applications of each scheme. It is imperative that the sampling frame and duration of recording provide a sufficient number of representative reports.

With interval-contingent recording, researchers must first determine the necessary number and timing of reports. Intervals should represent meaningful

time units, which depends on the target behavior's natural cycle, as well as the manner in which participants segment their activity. Intervals should be long enough to allow variation from one interval to the next, but not so long that successive fluctuations are masked or that forgetting or retrospection bias is likely. Intervals should be regular, with records scheduled at the same approximate point within each interval. For many behaviors, day's end meets these criteria.

The logic of signal contingency mandates a random schedule, although normal waking hours are often divided into fixed blocks, each with one random signal. The frequency and distribution of target activities and states in everyday life dictate the number and scheduling of blocks. Too few signals may bypass important occurrences, because only the moment of signal is described; too many signals may create excessive burden without incremental information yield. Typical ESM studies use 8–12 signals per day for 1–2 weeks. Delespaul (1992) provides a thorough discussion of various signaling plans.

Event-contingent sampling is predicated on unambiguous definition of the events to be recorded. Criteria should not be so inclusive as to overburden participants (and thereby invite sloppiness or noncompliance), but should be broad enough to include all instances of the target event. Because event-contingent sampling is commonly used to probe subtypes of general categories, some of which may be rare, data collection should continue as long as needed to obtain a reasonable number of each subtype. This period is likely to vary from one research topic to another. Social participation studies, for example, may require no more than one to two weeks, whereas studies of experience with prejudice might take considerably longer. In some cases, selective sampling schemes (e.g., every third meal or supervisor-supervisee interactions lasting 20 minutes or longer) may minimize participant burden while still yielding representative data.

All three formats require foreknowledge of the natural incidence and distribution of target events and states. Because research depends on spontaneous rather than manipulated events, data collection should be planned not only with base rates in mind, but also the frequency of cross-categorized activities, including contextual moderators of the process of interest. For example, one might be interested in the co-occurrence of alcohol consumption and sexual activity or social rejection – relatively rare conjunctions. Pilot data is therefore a must, to avoid discovering after the fact that the target phenomena have occurred insufficiently.

Because event sampling is designed to obtain representative data, it is best to avoid unusual circumstances. We would not study social interaction during final examinations, major holidays, or honeymoons, for example (unless these were the research focus). Nonetheless, choosing nonrandom times may inadvertently confound research. Enjoyable social interaction occurs more often on weekends than weekdays, for example, and depressed affect is more common in winter than in summer. These factors can be controlled by using multiple waves of assessment. Of course, even the best plans may be affected by unpredictable events. When such events are sample-wide (e.g., natural disasters), researchers may have stumbled onto an entirely new research focus. (For example, Cohn, Mehl, and Pennebaker [2004] examined linguistic markers of psychological functioning before and after the September 11 terrorist attack.)

Format of Administration

In the early days, everyday experience research used paper-and-pencil administration, often with convenient printed booklets, relying on pagers or programmable wristwatches to signal participants when a response was needed. Prearranged regular telephone interviews were also useful (Almeida, 1997), especially when prompts or interaction with an interviewer were deemed valuable. As technology advanced, more sophisticated alternatives became available, such as palmtop computers, programmed both to deliver prompts according to the research protocol and to record responses, and dedicated websites, to which participants would log in whenever a response was needed. More recent technological advances include specialized devices for administering protocols, such as the DIARYPro (www.invivodata.com) and the PsyMate (www.psymate.eu). Smartphones can also be used with programs that are directly installed (as of this writing, we are aware of several such programs that are being developed, but no well-tested programs are available for distribution) or as browsers to connect participants to data-collecting websites (see Maniaci and Rogge, Chapter 17 in this volume). Other, more specialized devices include the Electronically Activated Recorder for recording the auditory environment (Mehl & Robbins, 2012), devices that record various aspects of the context, such as location or proximity to other persons (Intille, 2012; Miller, 2012), and devices for making ambulatory recordings of psychophysiological processes (Wilhelm, Grossman, & Müller, 2012). Newer advances on the horizon include telemetric monitoring: wearable devices that “record behavioral, physiological, and environmental data from multiple sensors worn on the body

or embedded in the environment” (Goodwin, [2012](#), p. 251).

Computerized devices offer several major advantages over paper-and-pencil methods. One oft-cited benefit is the ability to verify the time of compliance with scheduled reports. Some investigators have questioned whether participants actually do complete reports at the time that they say they do or whether they practice *hoarding* (providing multiple reports from memory at a single moment; e.g., Broderick, Schwartz, Shiffman, Hufford, & Stone, [2003](#)). Other advantages include the ability to time-link self-reports with other data, such as ambulatory physiology or auditory recordings; to use branched protocols (asking different questions as a function of prior responses); to record response latency or duration; to randomize question order; and to upload data automatically and instantaneously, obviating data entry and allowing researchers to monitor compliance or data quality immediately rather than waiting until the end of data collection, when it is too late to intervene. On the other hand, paper-and-pencil methods also have advantages: they are inexpensive and easy to administer; participants often find them more convenient, in that no electronic interface is needed; and more diverse samples are possible, because computer literacy is not required. As for the problem of hoarding, participants can be given inexpensive time-stamping devices that cannot be altered.

Although the reasons for using electronic methods to collect self-report data are clear, there is no compelling evidence that either electronic or paper diaries produces better-quality data or results. Green, Rafaeli, Bolger, Shrout, and Reis ([2006](#)) compared electronic and paper diaries in three studies, concluding that they “were equivalent psychometrically and in the pattern of findings” (p. 87). Thus, the choice among these methods of administration should depend on each particular research setting and question. This conclusion notwithstanding, it seems inevitable that as portable technologies become more and more accessible, convenient, familiar, and inexpensive, electronic recording will become the de facto standard .

Designing the Instrument

Specific content depends, of course, on substantive interests. In hypothesis-driven studies, items should focus on the effect in question and its putative mediating mechanism, as well as alternative explanations and potential moderators. In exploratory research, items should be diverse, reflecting the full range and variability of a phenomenon, contextual factors, and expected or potential covariates. Information about likely predictors and consequences is

often desirable. Instruments are usually designed expressly for every study, although specific items are best selected on the basis of prior research.

Three kinds of self-report items are popular: open-ended questions, fixed-format rating scales, and checklists. Open-ended questions ask respondents to describe activity in their own words (e.g., “What were you thinking when the signal occurred?”). These responses are descriptively rich, but must be content-analyzed. The frequencies of conceptually relevant word categories can be examined using software such as the Linguistic Inquiry and Word Count program (Pennebaker & Francis, 1996). Open-ended items may be risky for hypothesis testing, in that laypersons’ spontaneous wordings may correspond imprecisely to hypothesized dimensions or features.

Traditional rating scales are used to quantify the degree to which particular qualities are present (e.g., mood, perceived control, intimacy). Likert-type and other rating scales facilitate quantitative analysis. Because the same questions are answered repetitively within a brief timespan, care must be taken to avoid reactivity and response-style artifacts (e.g., using the same scale values repeatedly). Stone *et al.* (1993) discuss methods for distinguishing internal consistency from response-style artifacts.

Checklists are used primarily for indicating which events within a series did and did not occur during a given interval, and then (possibly) rating some aspect of them. For example, checklists of stressful life events or of physical symptoms are common. Checklists require extensive development. Because they must be reasonably exhaustive without overwhelming respondents, they are usually limited to relatively common events, a practice that may omit rare or idiosyncratic but significant occurrences. (An open-ended “other event” category can be included.) Event definitions should be unambiguous and mutually exclusive, so that the same event is not tallied in separate categories at different times or by different respondents. A clear, standard format is desirable.

We have sometimes noticed confusion regarding the distinction between objective and subjective content. Whereas appraisals and evaluations are clearly subjective, other information, such as whether an interaction took place, in principle can be recorded objectively. Nevertheless, even such variables can have a subjective component. Respondents may fail to follow instructions or they may differ in perceiving whether or not a given event meets reporting criteria. For example, spouses may disagree about whether a conversation has been supportive, and some events are considered stressful by some persons but not by others. Gable, Reis, and Downey (2003) found that heterosexual dating

couples disagreed approximately 27% of the time about whether or not a supportive behavior had occurred on a given day. Because everyday experience studies are intended to maximize accuracy, protocols can and should minimize subjectivity whenever possible. In our opinion, objectivity is a matter of degree, and although everyday experience data generally come closer to that standard than other self-report methods do, the possibility of subjectivity should not be ignored.

The format of a protocol depends on several factors. Length and complexity should depend on participants' ability and motivation. Because excessive length inevitably degrades response quality, it is better to err on the side of brevity. We have found that once-daily reports should not exceed 15 minutes, and that more-than-daily reports should not exceed 5 minutes. Because of these limits, it is often not possible to include well-validated scales in their entirety, and it is common to measure key constructs with just a few items. Researchers must carefully balance the benefits of increased reliability provided by longer scales with the costs of increased participant burden (and resulting effects on attention and compliance). Presentation format also can make a big difference. A well-organized form that locates related items together with judicious use of formatting (e.g., page-scrolling, bolding, highlighting, and varied fonts) simplifies responding and improves data quality. For paper diaries, pocket-sized booklets are convenient and encourage timeliness. Questions presented on a digital device should be organized for clarity and simplicity, particularly if respondents will not be able to scroll backward. If protocols are administered on smartphones or the Internet, it is important to be sure the formatting works well for all relevant devices (desktop computers, laptops, tablets, etc; see Maniaci & Rogge, Chapter 17 in this volume).

Two final suggestions merit consideration. First, it is often useful to include general items about the topic being studied, both for descriptive purposes and to enhance the cover story. For example, in studying daily workload stress, we might ask how many hours the participant had worked that day and include an open-ended question about their activities, even if this material was not central to the research purpose. Second, signal-contingent and interval-contingent diaries should ask if anything important has happened since the prior report, which may clarify unusual circumstances .

Participant Issues

Event-sampling research inherently depends on participants' ability and

willingness to comply with instructions. Diaries are usually burdensome and at least somewhat intrusive, and ambulatory devices are often uncomfortable, necessitating care in participant relations. We always make the participant's task clear early on to minimize dropouts, as later dropouts compromise sampling and are costly (see Mazza & Enders, Chapter 24 in this volume). Incentives should be commensurate with burden: not so large as to attract freeloaders but not so small as to only attract research-eager volunteers. We find it helpful to highlight participants' role as collaborators in the "descriptive geography" of our work – that is, in developing a comprehensive factual map of everyday activity. (For example, we sometimes recruit participants by asking them to think about questions like, "How much time per day do people spend socializing?") Aside from diverting attention from theoretical aims, an intriguing rationale may enhance compliance. Protecting confidentiality, and ensuring that participants are aware of safeguards, is essential, especially when diaries ask about potentially embarrassing or compromising behaviors or mental states, or about illegal behaviors (e.g., domestic violence, drug use). In couple studies, it is vital that partners not have access to each others' records.

Detailed training to explain the protocol and item content is important. Some studies include one or two training days, whose data are excluded, so that questions and ambiguities can be resolved. Similarly, training can identify technical problems with electronic data collection (e.g., emails that are blocked by spam filters, smartphones that are out of contact). For paper diaries, to enhance and verify compliance, it is desirable to collect completed records and distribute new ones as often as possible, but certainly no less than every few days. Digital records should be monitored continuously. Debriefing interviews at the conclusion of record-keeping are useful for detecting problems such as noncompliance or misunderstanding of terms.

Documenting compliance is desirable and sometimes essential. The only ironclad assurances are computer-generated (or other nonalterable digital device) time stamps or interview-based administration. Recent evidence suggests that partial compliance may be a significant problem. Participants sometimes provide multiple reports at a single moment (Broderick et al., 2003). Litt, Cooney, and Morse (1998) studied drinking urges and behavior for 21 days with a combined signal-and event-contingent protocol. After extensive probing, 70% of their participants admitted delaying or omitting records at some time . Gable and Reis (1999) used computerized data collection to unobtrusively monitor the time at which daily activity records were completed. More than two-thirds of their sample recorded at least two days simultaneously, and 40% recorded four or

more days at once. With electronic devices, surveys can be programmed to allow responses only during predetermined times (e.g., a daily diary can be made inaccessible the following morning).

For some topics, partial compliance probably creates minimal distortion. People seem unlikely to forget that they were paid yesterday or that they spent the day alone. In other cases, particularly when the delay is long and the possibility of retrospection bias is substantial, meaningful inaccuracy may be introduced. When timeliness matters (as it does by definition with signal-contingent methods), researchers should monitor and promote compliance, such as by telephone calls, reminder emails, electronic recording, daily collection of materials, and added incentives for timeliness. (We have found lottery tickets given for each timely recording to be an effective incentive with student participants.) Excluding nontimely responses or full cases that exceed a predetermined noncompliance threshold is suboptimal, in our view. One reason is that delayed recording is unlikely to be random; it occurs most often when participants wish not to be interrupted. Furthermore, false negatives are more likely than false positives – omitting an event, signal, or interval rather than adding a nonexistent one.² The nature and extent to which these and other systematic distortions affect findings remains to be established.

Does Record-Keeping Alter Experience?

Because ecological validity is a prime justification for daily event studies, researchers sometimes worry about reactivity – that the process of record-keeping itself may alter the experiences being monitored. Anticipation of the need to describe an event, unavoidable with repetitive recording, may affect people's behavior choices or their experience of an event, especially if the events to be recorded activate impression management concerns. Even more pointedly, systematic attention to natural occurrences may foster unaccustomed levels of self-monitoring and new insights about the self (Bolger et al., 2003), modifying self-perceptions and perhaps leading to behavior change. For example, realizing that one has described a full day's social contacts as superficial may motivate intimacy strivings. Relatively few studies have directly investigated whether record-keeping alters experience, and those that have been conducted are inconclusive: A recent review by Barta, Tennen, and Litt (2012) included examples showing, in their words, no evidence, modest evidence, and reasonable evidence of reactivity effects. Unfortunately, even those studies that do show some evidence of reactivity effects did not address which procedural aspects are

responsible.

Reactivity is often indistinguishable from the general problem of response decay, in which repeated assessments alter participants' reports, such as when boredom leads to less thoughtful responding or to systematic differences in the frequency or nature of reports. One cause of response decay is participant burnout, which can occur when protocols are overly burdensome or a study runs too long. Diagnostic signs include diminished event frequency without apparent justification, more errors and missing data, and ratings stereotypy (e.g., logically inconsistent answers; changes in mean levels or variance across time, items, or events; and increased use of scale anchors or midpoints). Bolger and Laurenceau (in press) and Stone, Kessler, and Haythornthwaite (1991) describe statistical methods for addressing these problems.

The impact of repeated self-recording on behavior and subjective experience warrants study both as a methodological problem and as a substantive phenomenon. Substantively, event-sampling designs may allow researchers to investigate processes by which self-monitoring affects self-perception and behavior. For example, Conner and Reid (2011) asked participants to report their momentary happiness over two weeks for one, three, or six times each day. The more intensive protocol was associated with lower reports of momentary happiness over time among participants higher in neuroticism and depressive symptoms, but increased reports of happiness among participants lower in neuroticism and depressive symptoms. Methodologically, no clear conclusions can be drawn at this time about the extent to which reactivity may influence the findings of event-sampling research. Researchers must be aware of the possibility that their procedures may alter the behavior being studied – a legitimate concern given the rationale of studying behavior in its natural context. Research is also needed to determine how to design event-sampling protocols to minimize potential problems. For now, we conclude that it is best to identify such problems in pilot-testing and revise procedures accordingly.

Data Analytic Strategies and Considerations

The unique nature of event-sampling data requires special consideration for data analysis. The simplest issue – managing the sheer mass of data that repeated self-recording provides – has become increasingly less burdensome over time. Other issues, such as nesting, serial dependence, and imbalance in the number and variance of data points, are more complex. However, the commitment of

time and effort that event-sampling research requires makes maximization of information yield a necessity. We therefore recommend that researchers have a firm grasp of data analytic strategies and practices in the earliest stages of design and planning so that their study designs and data analytic strategies mirror their questions of interest. In this section we review design and conceptualization issues relevant to data analysis. Detailed discussions of technical and computational matters can be found in other sources, including Schoemann, Rhemtulla, and Little (Chapter 21 in this volume) and Bolger and Laurenceau (2013), as well as in Mehl and Conner's (2012) comprehensive edited volume.

The most typical daily experience designs yield two levels of data: measurements, nested within individuals. From this structure three general types of questions can be posed. One question type focuses on the higher unit of analysis – differences between individuals – and incorporates between-person effects and hypotheses (e.g., Do men and women differ in their average daily health behaviors?). Another type of question considers the lower unit of analysis – variation within individuals – and incorporates within-person effects and hypotheses (e.g., Do health behaviors vary by day of the week? Is daily stress associated with daily health behaviors?). A third question type examines interactions between the upper and lower levels – whether within-person variations differ as a function of between-person factors (e.g., Is variation in health behaviors across the week different for men and women? Or, is the association between daily stress and health behaviors different for men and women?). Designs with more than two levels are also feasible, such as when measurements are nested within employees who are in turn nested within workgroups. In this three-level data set, the highest level of analysis entails between-groups questions (e.g., do large workgroups differ from small workgroups on daily job satisfaction?) in addition to between-person, within-person, and cross-level questions.

Researchers should be careful to frame questions at the appropriate level of analysis. As discussed earlier, within-person hypotheses are sometimes framed in between-person terms for simplicity, leading to conceptual, and occasionally statistical, misspecification (Gable & Reis, 1999). Although event sampling is often used to address purely between-person hypotheses, within-person and cross-level questions represent more compelling models for everyday experience research. Another consideration is that event-sampling data frequently require a choice between relatively simple, heuristically clear techniques and more elegant but also more complex statistical methods. Although we subscribe to the maxim of simplicity, we firmly believe that statistical advances inevitably enhance the

quality of insights that data can produce. We therefore encourage researchers to seek an optimal synthesis of heuristic simplicity, time-tested practices, and the enhanced precision of newer methods .

Aggregation and Composites

The simplest way to analyze event data is to calculate numerical composites across all of an individual's records,³ which corresponds to the highest level of analysis (individuals) mentioned earlier. For example, from 14 days of data, one might compute the total number of social interactions, trips to the gym, or times participants felt discriminated against; mean levels of stress, perceived rejection, or attempted suppression of emotion; variability (e.g., the within-person standard deviation) of mood, self-esteem, or calories consumed; or the proportion of all social contacts that involved one's in-group. Using standard statistical tests such as regression and ANOVA, these composites can be related to predictors, such as personality or demographic or situational variables, or to outcomes, such as health, well-being, or productivity.

This strategy builds on the well-known advantages of composites for maximizing reliability (Epstein, 1983) and is essential for descriptive data. There are several liabilities, however, beginning with the loss of potentially valuable information at the within-person level, which is particularly unfortunate, given the great effort that event-sampling research entails. Another drawback concerns reliability: To the extent that the number of records per person varies or the within-person variance differs from one person to another, the resulting statistics may improperly estimate standard errors and significance levels.⁴ Per-person record counts are likely to vary sharply in event-contingent sampling, an issue that also applies to time-contingent and signal-contingent sampling, in which the relative frequency of subcategories may vary (e.g., one person signaled 100 times may be with her spouse 10 times and with others 90 times, whereas a second person may display the inverse ratio).

There are good reasons to rely on the computational and heuristic simplicity of composites when strictly between-persons questions are of interest. Nevertheless, possibilities for misleading conclusions should always be explored. We advocate a two-tiered approach, in which aggregation is used for descriptive purposes – as noted earlier, knowing how often, to what degree, or in what context a construct occurs is an often overlooked but necessary step in understanding a phenomenon – before continuing onto multilevel analyses that take nesting into proper account .

The Logic of Multilevel Modeling of Everyday Experience Data

Suppose a research team measured momentary stress and selfregulatory resources in a 14-day interval-contingent daily diary study. As described in the preceding section, the researchers could address a strictly between-persons question by calculating mean levels of stress or selfregulatory resources and determining if these means are associated with some personal demographic or dispositional variable (e.g., age or neuroticism). Because each daily record includes participants' reports of the two variables of interest, across the study one can calculate the association (i.e., slope) of stress to selfregulation for each person. This slope would vary to the extent that some people show a strong positive association and others a weak or even negative association. Calculating the average slope across participants would address a strictly within-persons question and estimate the average association between daily stress and selfregulatory resources in the population. In addition, variations between people in the magnitude or direction of this within-person association may be meaningful: The person-by-person slopes might be correlated with demographic or dispositional variables. This is the cross-level interaction question.

These analyses could be conducted with traditional linear regression or ANOVA (e.g., by removing individual differences with dummy codes or by treating each person's slope as a variable). However, such analyses are problematic for several reasons, including that frequency and variability differences from one person to another may bias significance tests; misspecified error terms and incorrect degrees of freedom because of treating random effects as fixed; and the inability to deal with serial autocorrelation (the tendency of scores obtained close in time to be more highly correlated than those obtained further apart in time).

Multilevel modeling (also called hierarchically nested models, hierarchical linear models, or latent growth-curve modeling) offers several key advantages for the analysis of everyday experience data. Multilevel models allow simultaneous estimation of between-and within-persons effects and their interaction; they readily handle multiple continuous predictors with an unbalanced number of cases per person (including missing data; see Mazza & Enders, Chapter 24 in this volume); and they simplify treating variables as random rather than fixed effects. Also, importantly, multilevel modeling is based on maximum likelihood estimation, which, although computationally complex,

is more precise and efficient in many data situations than least squares estimation. (See Schoemann et al., Chapter 21 in this volume, for a more in-depth discussion of multilevel modeling.) Daily experience data can also be analyzed with structural equation modeling (Eid, Courvoisier, & Lischetzke, 2012). However, the conceptual logic of multilevel modeling follows straightforwardly from regression, and because the concepts of slopes and intercepts are already familiar to most social-personality psychologists, we focus on them here.

Revisiting our 14-day study of stress and selfregulatory resources, let us assume that a dispositional measure of self-esteem is also available, which may moderate the stress-resource relationship. The upper-level unit of analysis here is the person, whereas the lower-level unit, nested within persons, is the day. For each person, daily regulatory resources are estimated as:

$$y_{ij} = b_{0j} + b_{1j}X_{ij} + e_{ij}, \quad (15.1)$$

where Y_{ij} refers to each individual's regulatory resources on a given day (i.e., the i th day for the j th participant), b_{0j} refers to that individual's average resources across all 14 days⁵, X_{ij} is the individual's stress rating for that day, b_{1j} is the regression coefficient indicating the degree of change in regulatory resources produced by a one-unit change in stress on a given day, and e_{ij} is error. The innovative part of multilevel modeling is in estimating b_{0j} and b_{1j} (intercepts and slopes, respectively) for each individual participant. The constant term (or intercept) for each individual, b_{0j} , is represented as:

$$b_{0j} = a_0 + a_1Z_i + u_{0i}, \quad (15.2)$$

where a_0 refers to the sample-wide mean selfregulatory resource score, Z_i is the individual's self-esteem rating, a_1 is the coefficient indicating the degree of change in individual's mean regulatory resources produced by a one-unit change in self-esteem, and u_{0i} is error. The slope representing the effect of daily stress on daily selfregulatory resources for each individual, b_{1j} , is:

$$b_{1j} = c_0 + c_1Z_i + u_{1i}, \quad (15.3)$$

where c_0 represents the average effect of stress on regulatory resources in the sample, Z_i is self-esteem, c_1 is the regression coefficient now indicating the degree of change in the stress-resources slope produced by a one-unit change in self-esteem, and u_{1i} is error. In other words, c_1 tells us whether the stress-resources relationship is moderated by self-esteem. If c_1 is trivial, the slope representing the effect of daily stress on daily selfregulatory resources will be similar when computed separately for individual participants. If c_1 is large, individual slopes will be larger or smaller than the sample average, depending on whether their self-esteem is higher or lower than the sample mean.

Examining Equations 15.1 and 15.2 shows that on an average stress day (i.e., $X_{ij} = 0$, after centering), an individual's predicted regulatory resource score is solely a function of dispositions and error. Similarly, Equations 15.1 and 15.3 reveal that for someone whose self-esteem is average (i.e., $Z_i = 0$), the magnitude of daily selfregulatory resource fluctuations mirrors the sample-wide average effect. Substituting Equations 15.2 and 15.3 into Equation 15.1 yields:

$$y_{ij} = a_0 + a_1 Z_i + c_0 X_{ij} + c_1 Z_i X_{ij} + u_{0i} + u_{1i} X_{ij} + e_{ij} \quad (15.4)$$

The second, third, and fourth terms on the right side of the equation are main effects for Z , X , and their interaction, e_{ij} is lower-level error term, u_{0i} and $u_{1i}X_{ij}$ represent the random effect for the intercept and slope, respectively.

We can now illustrate how these equations address questions typical to event-sampling research. The classic between-persons question is, “Is self-esteem related to average selfregulatory resources?” and is quantified by the intercept term a_1 in Equation 15.2. The strictly within-persons question – “On days in which stress levels are relatively high, are selfregulatory resources lower?” – is assessed by the sample-average slope, c_0 , in Equation 15.3. This effect is independent of dispositional differences in resources or stress. The third and final question asks about the interaction: “Does the stress–resource relationship vary in magnitude as a function of self-esteem?” The term c_1 in Equation 15.3 provides the appropriate test. Because interpretation of interactions can be tricky, Equation 15.3 should be recomputed for meaningfully informative levels of self-esteem (e.g., one standard deviation above and below the mean). Significant interactions will be evident in nonparallel slopes .

Complex Multilevel Models

Multilevel models can be expanded to incorporate multiple predictors at each level of analysis, for example to compare the relative contribution of stress and social support to regulatory resources. In addition, multiple upper-level predictors can also be included, such as examining how sex and self-esteem moderate intercepts and slopes.

Of growing appeal to daily experience researchers is the ability of multilevel models to accommodate additional levels of nesting (although in current practice, models with more than three levels are difficult to analyze). If our stress resources example had been conducted with subjects who participate in different support groups, days would be nested within individuals, who would in turn be nested within groups. A particularly common example of a three-level model involves romantic couples – days would be nested within persons, who are in turn nested within dyads (see Kenny & Kashy, Chapter 22 in this volume).⁶ Thus, the highest-level unit of nesting need not be the person; any grouping in which scores are nested (correlated) is appropriate.

In some instances it may be desirable to aggregate across one level of nesting prior to analysis. For example, one might calculate composites for all interactions of a given sort, with particular partners, or on certain days of the week. Although this procedure likely violates the equal-reliability assumption discussed earlier, implications of this violation are unknown. The great advantage of pre-aggregation is to simplify data management and to reduce some of the inherent (and perhaps conceptually irrelevant) variability in everyday events. Collapsing across minor variations within a category may also provide more meaningful indicators of a construct.

Another level of complexity concerns mediation (Judd et al., Chapter 25 in this volume). There is insufficient space in this chapter to discuss these complexities, but it bears noting that mediation is handled somewhat differently at each level of a multilevel model. Interested readers are referred to Bauer, Preacher, and Gil (2006), Kenny, Korchmaros, and Bolger (2003), and Preacher, Zyphur, and Zhang (2010). We anticipate that such analyses will soon be commonplace in the social-personality literature.

Multilevel modeling programs (e.g., HLM, M-Plus, SAS PROC MIX, SPSS Mixed) were developed to capitalize on ML estimation, especially with unbalanced data sets, but they have come to offer another benefit – the ability to maximize information retrieval from large, hierarchically nested data sets. In the

past decade, multilevel modeling has become the gold standard for analyzing everyday experience data. As with all statistically complex methods, it is critical to not lose sight of the conceptual foreground. This means thinking about slopes and intercepts, the two elementary constructs. It also means not abandoning simpler statistics. Researchers must be prepared to inspect data thoroughly and to rely on multiple analytic strategies before “going to the bank” with findings. Distributions should be scrutinized carefully, particularly for outliers (univariate and multivariate), heteroscedasticity, and correlated errors. Models should be based on substantive theory, taking full advantage of the ability of daily experience data to address propositions from between-person, within-person, and temporal perspectives. Studies should be designed, and sample sizes chosen, to provide appropriate levels of power at each level (Bolger, Stadler, & Laurenceau, 2012).⁷ Appeasing the competing muses of statistical precision and heuristic clarity is a fine art that all everyday experience data analysts must master.

Analysis of Temporal Patterns

Although spurious causation cannot be discounted in the absence of experimental manipulation, temporal precedence in sequentially structured data can support certain hypotheses while ruling out others (West, Cham, & Liu, [Chapter 4](#) in this volume). This type of test is commonly conducted by examining lagged effects – variable Y at time t predicted from variable X at time $t - 1$, controlling for variable Y at time $t - 1$. Synchronous correlations – variable Y at time t predicted from variable X at time t , controlling for variable Y at time $t - 1$ – provide much less persuasive evidence of a causal effect. Lagged tests of causation are especially impressive because of the repetitive nature of event-sampling data – in a 14-day daily diary study, the lagged effect is tested 13 times. For example, Downey *et al.* (1998) collected daily ratings of perceived rejection, conflict, and relationship satisfaction from romantic partners for four weeks. Following days in which they felt rejected, rejection-sensitive women reported significantly more conflict and their partners reported lesser satisfaction. The lagged effect allowed Downey and colleagues to demonstrate temporal precedence, supporting a causal interpretation, while simultaneously minimizing the plausibility of the reverse causal path (i.e., that conflict engenders feelings of rejection).

More generally, variations in how the passage of time is experienced and how circumstances at a given moment may influence subsequent behavior embody

potentially fruitful questions for understanding social life. Sequential data provide substantive opportunities for studying cyclical patterns, as well as for causal analysis. Disregarding sequence in diary data therefore represents underutilization of a valuable resource.

Some phenomena are likely to be revealed in sequential processes. For example, Margolin, Christensen, and John (1996) obtained daily telephone reports of family conflict for two weeks, subdivided into mornings, afternoons, and evenings. Distressed and nondistressed families were differentiated by continuance of tension from one day to the same time the next day, but not at shorter intervals. In other words, the duration of a conflictual atmosphere may be the hallmark of familial distress.

Spectral analysis, a general class of methods for detecting cycles and rhythms in sequential data (Larsen, 1990), is rarely applied in social-personality psychology, despite the prevalence of cycles in our phenomena. One well-known cycle is the day, characterized by regularities in activity schedules and diurnal rhythms in internal states such as mood, fatigue, and attentiveness. People also experience weekly cycles, which in one study accounted for 40% of the variance in daily mood (Larsen & Kasimatis, 1990). Mood tends to be more positive and less negative on weekends than weekdays, although there is disagreement about whether mood varies reliably from Monday to Friday (Stone et al., 1985). Larsen (1987) also identified a monthly mood cycle. Other relevant temporal cycles include seasons and academic semesters, in which predictable markers (e.g., weather, stress, study patterns) may be influential. A key focus for social-personality psychologists would be to identify the mechanisms that underlie such patterns.

There is another, more statistical rationale for attending to sequence: the possibility of serial dependency, which can produce spurious correlations between observations adjacent in time. There are several reasons for this, the most common being that error terms in two consecutive observations are unlikely to be independent (as standard analyses assume). For example, mood ratings taken at 2 PM and 4 PM on the same day may be influenced by the same outside factors (e.g., weather, office environment); thus error in Time 1 mood ratings is correlated with error in Time 2 mood. All other things being equal, the closer in time or similarity in context two observations are, the larger these effects are likely to be. With many closely spaced observations, such as in signal-contingent recording, serial dependency may be substantial. Correlated residuals, as this problem is also called, may bias standard errors (and hence

significance tests) and may distort correlations among sequential within-person variables. Because of this, we recommend that researchers seeking to draw causal or quasi-causal conclusions from temporal patterns within everyday experience data sets learn more about these methods. Bolger and Laurenceau (in press) discuss these issues more fully, and West and Hepworth (1991) offer an excellent introduction to strategies for identifying and contending with serial dependency.⁸ Readers seeking a more in-depth presentation are referred to Gottman and Roy (1990) or Fitzmaurice, Laird and Ware (2011).

Time-series analyses are rare in psychology but common in other social sciences. We anticipate that these analyses will become increasingly familiar to social-personality psychologists, given their usefulness for analyzing the extensive streams of continuous data produced by ambulatory assessment of physiology and activity. Emerging technologies make possible increasingly sophisticated and comprehensive data sets in this regard (Miller, 2012), and researchers will want to avail themselves of the statistical methods that can best derive conceptual insights from them. Ebner-Priemer and Trull (2012) provide a useful introduction to the application of temporal analysis in everyday experience studies .

Integrating Everyday Experience Methods in Programmatic Research

In [Chapter 2](#) of this *Handbook*, Brewer and Crano argued that construct validity is a property of research programs, not of individual studies. If so, the value of event sampling for social-personality psychology should be appraised not in isolation but rather in terms of how it complements other methods. Every method has benefits and drawbacks, insights that it can and cannot impart. Methodological triangulation – embracing diverse strategies and procedures – is the best way to prevent theoretical insights from being method-bound. Of course, methods ought not to be chosen willy-nilly; they should complement one another in addressing conceptual concerns and methodological shortcomings.

Webb, Campbell, Schwartz, and Sechrest (1966) introduced the idea of *multiple operationalism* to suggest that because all methods have strengths and weaknesses, the most useful perspectives on a phenomenon are provided by methods that complement each other's limitations. All empirical tests have substantive and methodological components that contribute to their outcomes. Determining the degree to which results should be attributed to each component

is fundamentally impossible, as Houts, Cook, and Shadish (1986) noted, without corroboration from “aggressively sought alternative explanations” (p. 56), including both methodological and conceptual explanations. They therefore argued for “critical multiplism,” which substitutes reliance on diverse alternative methods for (perhaps naïve) faith in a single critical test or paradigm.

Daily experience studies represent a valuable addition to the social-personality psychology toolbox in this regard. For example, one might explore stereotype threat processes, as identified and interpreted in laboratory experiments, with an event-contingent sampling scheme. Corroborated principles can be held in greater confidence; propositions not confirmed indicate the need for further theorizing. Certain effects may occur only in the presence of particular moderating conditions, for example. Or, a process may be confounded with another process or procedural detail, implying incorrect specification of the underlying theory. Still other findings may suggest clarifications or extensions of a theory, leading to further experiments. In any or all of these eventualities, the synthesis of laboratory and everyday experience approaches can be expected to yield fascinating insights and more compelling theories.

A research program might also start with event sampling. As discussed earlier, everyday experience studies can describe the prevalence of phenomena and their predictors, covariates, and consequences. Establishing the scope and nature of a phenomenon, as well as its natural context, is central to basic science, as the biological and physical sciences have historically shown, and we are confident that our discipline would benefit from such data. John Bowlby (1988, pp. 40–41) implied as much:

[The wise researcher] will concentrate attention on a limited aspect of a limited problem. If in making his selection he proves sagacious, or simply lucky, he may not only elucidate the problem selected but also develop ideas applicable to a broader range. If his selection proves unwise or unlucky he may merely end up knowing more and more about less and less.

We do not mean to suggest that the benefits of everyday experience research reside only, or even primarily, in replicating existing findings or in identifying phenomena to be taken apart in the laboratory. At the outset of this chapter, we discussed direct theoretical uses of everyday experience studies, including tests of hypotheses in natural contexts, comparison of competing predictions and models, identification of moderators, evaluation of alternative explanations, and separation of within-person from dispositional processes. Many of the specific

examples cited earlier demonstrate the benefits of theory testing with everyday experience data.

To be sure, the sine qua non of theoretical specification in social psychology is cause-and-effect. Internally valid causal inference requires three conditions: correlation, temporal precedence, and nonspuriousness. This last factor is most definitively established with experimenter-manipulated conditions, carefully controlled contexts, and random assignment. Causal relations are, and should remain, a central focus of psychological research, but, as described earlier and elsewhere in this *Handbook*, laboratory experiments have limitations. Everyday experience studies provide an alternative perspective, complementing the laboratory not only in method but also in the kind of behavior examined and in the contexts in which they are evidenced. This is the logic behind multiple operationalism.

Likewise, momentary reports offer a useful counterpart to global self-reports. Although many phenomena are inherently global and subjective (e.g., attitudes, self-perceptions), for other constructs such measures inevitably beg the question of whether actual experience or after-the-fact reconstruction is being assessed. Comparison of findings from momentary and retrospective accounts can shed light on these important processes.

Lest we mislead readers, we do not recommend daily experience research universally. At least five situations seem ill-suited to this form of research: when an effect is unlikely to be evident without careful control of other, simultaneous processes; when, in fine-tuning a theory, it is necessary to create unusual, rare, or even implausible conditions; when ruling out spurious causation is paramount; when global, general impressions are the theoretical focus; and when participants are unable or unlikely to comply with sampling protocols.

On the other hand, event sampling seems well suited to at least five circumstances: when it is desirable to observe phenomena in natural, voluntary, and spontaneous contexts; when retrospection and certain other biases are likely to produce misleading accounts; when relevant conditions cannot be created ethically or impactfully in the laboratory; when within-person and temporal patterns are likely to be revealing; and when ecological validity is foremost.

In sum, validity, defined in the broadest possible sense, requires methodological triangulation. Our general impression is that social-personality psychologists often pursue methodological variety with relatively minor variations on the same theme. Everyday experience methods, in conjunction with

laboratory and global self-report strategies, offer a substantial alternative with which to enhance the validity of a research program .

Concluding Comments

Social psychology's special niche in the behavioral and social sciences concerns the impact of the social environment on behavior. Most social psychologists take as an article of faith the importance of understanding the situational context in which behavior occurs. Reflected throughout the literature, this orientation is evident in the popular construct of Personality \times Situation interactionism – that dispositions are reflected in the situations that people select, in differential reactions to existing situations, and in the way that one person's behavior may alter the situation for others.

Yet too often research seems oblivious to this core principle. For example, many studies treat behavior as essentially acontextual or as a fixed quality of individuals (even if theorizing from a more dynamic contextual view). Measurement practices often seem to assume that behavior does not depend on the context of observation or the manner of assessment. Methodology is construed as a search for operations that yield significant effects, and self-reports are treated alternatively as *prima facie* accurate or as irreparably biased. Of course most researchers know otherwise. The fifth *Handbook of Social Psychology* has been published, but this is only the second edition of a methods handbook. An entire handbook devoted to methods suggests that the research process in social psychology is changing and that methodology has become a fundamentally more complex and diverse business than before.

Without doubt, social psychological theorizing is increasingly sophisticated, detailed, and differentiated. Whereas studies were once focused on uncovering core phenomena, the cumulative knowledge produced by decades of research now directs our attention to second-order questions: identifying moderators and boundary conditions for basic processes, verifying underlying mechanisms (mediating processes) for established phenomena, and determining the relevance and applicability of competing theories and predictions. The growing complexity of theories, and hence the most cutting-edge questions, inevitably requires more diverse methods, both to address more pointed questions and to compensate for the inescapable limitations of any single method. In short, the need for – but also the opportunities provided by – multiple operationalism is greater than ever.

Everyday experience studies offer one avenue for bridging some of the field's

more imposing gaps: between laboratory experiments and survey questionnaires, between one-shot observations and global retrospections, between theorizing about contextual variations and examining them empirically, between internal and ecological validity, between highly controlled situations and those encountered in natural activity, and between abstract theories and the details of ordinary life. Integrated with the many other methods represented in this *Handbook*, everyday experience methods can help the next generation of research fulfill our legacy of understanding how social behavior is embedded in situational contexts .

References

- Almeida, D. M. (1997). *National study of daily experiences: The MIDUS in-depth diary study. Daily inventory of stressful events: Expert coding manual*. Chicago: MacArthur Foundation Research Network on Successful Midlife Development.
- Asch, S. E. (1952). *Social psychology*. New York: Prentice Hall.
- Augustine, A. A., Mehl, M. R., & Larsen, R. J. (2011). A positivity bias in written and spoken English and its moderation by personality and gender. *Social Psychological and Personality Science*, 2, 508–515.
- Barta, W. D., Tennen, H., & Litt, M. D. (2012). Measurement reactivity in diary research. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 108–123). New York: Guilford Press.
- Bauer, D. J., Preacher, K. J., & Gil, K. M., (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, 11, 142–163.
- Beckmann, N., Wood, R. E., & Minbashian, A. (2010). It depends how you look at it: On the relationship between neuroticism and conscientiousness at the within-and the between-person levels of analysis. *Journal of Research in Personality*, 44, 593–601.
- Berkman, E. T., Falk, E. B., & Lieberman, M. D. (2011). In the trenches of real-world self-control: Neural correlates of breaking the link between craving and smoking. *Psychological Science*, 22, 498–506.
- Bernard, H. R., Killworth, P., Kronenfeld, D., & Sailer, L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of*

Anthropology, 13, 495–517.

- Birnbaum, G. E., Reis, H. T., Mikulincer, M., Gillath, O., & Orpaz, A. (2006). When sex is more than just sex: Attachment orientations, sexual experience, and relationship quality. *Journal of Personality and Social Psychology*, 91, 929–943.
- Blaney, P. H. (1986). Affect and memory: A review. *Psychological Bulletin*, 99, 229–246.
- Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, 54, 579–616.
- Bolger, N., DeLongis, A., Kessler R. C., & Schilling E. A., (1989). Effects of daily stress on negative mood. *Journal of Personality and Social Psychology*, 57, 808–818.
- Bolger, N., & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York: Guilford Press.
- Bolger, N., Sadler, G., & Laurenceau, J. P. (2012). Power analysis for intensive longitudinal studies. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 285–301). New York: Guilford Press.
- Bolger N., & Schilling E. A. (1991). Personality and the problems of everyday life: The role of neuroticism in exposure and reactivity to daily stressors. *Journal of Personality*, 59, 355–386.
- Bolger N., & Zuckerman, A. (1995). A framework for studying personality in the stress process. *Journal of Personality and Social Psychology*, 69, 890–902.
- Bowlby, J. (1988). *A secure base: Parent-child attachment and healthy human development*. New York: Basic Books.
- Brewin, C. R., Andrews, B., & Gotlib, I. H. (1993). Psychopathology and early experience: A reappraisal of retrospective reports. *Psychological Bulletin*, 113, 82–98.
- Broderick, J. E., Schwartz, J. E., Shiffman, S., Hufford, M. R., & Stone, A. A. (2003). Signaling does not adequately improve diary compliance. *Annals of*

Behavioral Medicine, 26, 139–148.

- Brown, K. W., & Moskowitz, D. S. (1997). Does unhappiness make you sick? The role of affect and neuroticism in the experience of common physical symptoms. *Journal of Personality and Social Psychology*, 72, 907–917.
- Campbell, J. D., Chew, B., & Scratchley, L. S. (1991). Cognitive and emotional reactions to daily events: The effects of self-esteem and self-complexity. *Journal of Personality*, 59, 473–505.
- Campbell, L., Simpson, J. A., Boldry, J. G., & Rubin, H. (2010). Trust, variability in relationship evaluations, and relationship processes. *Journal of Personality and Social Psychology*, 99, 14–31.
- Carlston, D. E., & Smith, E. R. (1996). Principles of mental representation. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 184–210). New York: Guilford Press.
- Carstensen, L. L., Pasupathi, M., Mayr, U., & Nesselroade, J. R., (2000). Emotional experience in everyday life across the adult life span. *Journal of Personality and Social Psychology*, 79, 644–655.
- Caspi, A., Bolger, N., & Eckenrode, J. (1987). Linking person and context in the daily stress process. *Journal of Personality and Social Psychology*, 52, 184–195.
- Clark, L. A., & Watson, D. (1988). Mood and the mundane: Relations between daily life events and self-reported mood. *Journal of Personality and Social Psychology*, 54, 296–308.
- Clark, M. S., Fitness, J., & Brissette, I. (2001). Understanding people's perceptions of relationships is crucial to understanding their emotional lives. In M. Hewstone & M. Brewer (Eds.), *Blackwell handbook of social psychology: Interpersonal processes* (Vol. 2, pp. 253–278). Oxford: Blackwell Publishers.
- Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic indicators of psychological change after September 11, 2001. *Psychological Science*, 15, 687–693.
- Conner, T., & Barrett, L. F. (2012). Trends in ambulatory self-report: The role of momentary experience in psychosomatic medicine, *Psychosomatic Medicine*, 74, 327–337.

- Conner, T. S., & Reid, K. (2011). *Intensive momentary reporting of happiness*. Manuscript under review.
- Conrath, D. W., Higgins, C. A., & McClean, R. J. (1983). A comparison of the reliability of questionnaire versus diary data. *Social Networks*, 5, 315–322.
- Csikszentmihalyi, M., & Larson, R. (1984). *Being adolescent*. New York: Basic Books.
- Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *Journal of Nervous and Mental Disease*, 175, 526–536.
- Csikszentmihalyi, M., Larson, R. W., & Prescott, S. (1977). The ecology of adolescent activity and experience. *Journal of Youth and Adolescence*, 6, 281–294.
- David, J. P., Green, P. J., Martin, R., & Suls, J. (1997). Differential roles of neuroticism, extraversion, and event desirability for mood in daily life: An integrative model of top-down and bottom-up influences. *Journal of Personality and Social Psychology*, 73, 149–159.
- Decastro, J. M., & Pearcey, S. M. (1995). Lunar rhythms of the meal and alcohol intake of humans. *Physiology & Behavior*, 57, 439–444.
- Delespaul, P. A. E. G. (1992). Technical note: Devices and time-sampling procedures. In M. W. de Vries (Ed.), *The experience of psychopathology: Investigating mental disorders in their natural settings* (pp. 363–373). Cambridge, MA: Cambridge University Press.
- Delespaul, P. A. E. G. (1995). *Assessing schizophrenia in daily life: The experience sampling method*. Maastricht, The Netherlands: International Institute for Psycho-social and Socio-ecological Research.
- Delongis, A., Folkman, S., & Lazarus, R. S. (1988). The impact of daily stress on health and mood: Psychological and social resources as mediators. *Journal of Personality and Social Psychology*, 54, 486–495.
- DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, 70, 979–995.
- Diamond, L. M., Hicks, A. M., & Otter-Henderson, K. D. (2008). Every time

you go away: Changes in affect, behavior, and physiology associated with travel-related separations from romantic partners. *Journal of Personality and Social Psychology*, 95, 385–403.

Diener, E. (1996). Traits can be powerful, but are not enough: Lessons from subjective well-being. *Journal of Research in Personality*, 30, 389–399.

Diener, E., Larsen, R. J., Levine, S., & Emmons, R. A. (1985). Intensity and frequency: Dimensions underlying positive and negative affect. *Journal of Personality and Social Psychology*, 48, 1253–1265.

Downey, G., Freitas, A. L., Michaelis, B., & Khouri, H. (1998). The self-fulfilling prophecy in close relationships: Rejection sensitivity and rejection by romantic partners. *Journal of Personality and Social Psychology*, 75, 545–560.

Duck, S., Rutt, D. J., Hurst, M. H., & Strejc, H. (1991). Some evident truths about conversations in everyday relationships: All communications are not created equal. *Human Communication Research*, 18, 228–267.

Ebner-Priemer, U. W., & Trull, T. J. (2012). Investigating temporal instability in psychological variables: Understanding the real world as time dependent. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 423–439). New York: Guilford Press.

Eid, M., Courvoisier, D. S., & Lischetzke, T. (2012). Structural equation modeling of ambulatory assessment data. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 384–406). New York: Guilford Press.

Eisenberger, N. I., Gable, S. L., & Lieberman, M. D. (2007). Functional magnetic resonance imaging responses relate to differences in real-world social experience. *Emotion*, 7, 745–754.

Emmons, R. A. (1991). Personal strivings, daily life events, and psychological and physical well-being. *Journal of Personality*, 59, 453–472.

Emmons, R. A., & King, L. A. (1988). Conflict among personal strivings: Immediate and long-term implications for psychological and physical well-being. *Journal of Personality and Social Psychology*, 54, 1040–1048.

Emmons, R. A., & King, L. A. (1989). Personal striving differentiation and affective reactivity. *Journal of Personality and Social Psychology*, 56, 478–

484.

- Epstein, D. H., Marrone, G. F., Heishman, S. J., Schmittner, J., & Preston, K. L. (2010). Tobacco, cocaine, and heroin: Craving and use during daily life. *Addictive Behaviors*, 35, 318–324.
- Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality*, 51, 360–392.
- Feldman-Barrett, L., & Pietromonaco, P. R. (1997). Accuracy of the five-factor model in predicting perceptions of daily social interactions. *Personality and Social Psychology Bulletin*, 23, 1173–1187.
- Feldman-Barrett, L., Robin, L., Pietromonaco, P. R., & Eyssell, K. M. (1998). Are women the more emotional sex? Evidence from emotional experience in social context. *Cognition & Emotion*, 12, 555–578.
- Finkel, E. J., DeWall, C. N., Slotter, E. B., McNulty, J. K. Pond, S. R., & Atkins, D. C. (2012). Using I³theory to clarify when dispositional aggressiveness predicts intimate partner violence perpetration. *Journal of Personality and Social Psychology*, 102, 533–549.
- Fiske, S. T., & Taylor, S. E. (2007). *Social cognition* (3rd ed.). New York: McGraw-Hill.
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.
- Fleeson, W. (2001). Toward a structure-and process-integrated view of personality traits as density distributions of states. *Journal of Personality and Social Psychology*, 80, 1011–1027.
- Fleeson, W., & Nofle, E. E. (2012). Personality research. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 525–538). New York: Guilford Press.
- Fraley, R. C., Vicary, A. M., Brumbaugh, C. C., & Roisman, G. I. (2011). Patterns of stability in adult attachment: An empirical test of two models of continuity and change. *Journal of Personality and Social Psychology*, 101, 974–992.
- Fujita, F., & Diener, E. (2005). Life satisfaction set point: Stability and change. *Journal of Personality and Social Psychology*, 88, 158–164.

- Funder, D. C. (1991). Global traits: A neo-Allportian approach to personality. *Psychological Science*, 2, 31–39.
- Gable, S. L., & Reis, H. T. (1999). Now and then, them and us, this and that: Studying relationships across time, partner, context, and person. *Personal Relationships*, 6, 415–432.
- Gable, S. L., Reis, H. T., & Downey, G. (2003). He said, she said: A quasi-signal detection analysis of daily interactions between close relationship partners. *Psychological Science*, 14, 100–105.
- Gable, S. L., Reis, H. T., & Elliot, A. J. (2000). Behavioral activation and inhibition in everyday life. *Journal of Personality and Social Psychology*, 78, 1135–1149.
- Gallo, L. C., Bogart, L. M., Vranceanu, A. M., & Matthews, K. A. (2005). Socioeconomic status, resources, psychological experiences, and emotional responses: A test of the reserve capacity model. *Journal of Personality and Social Psychology*, 88, 386–399.
- Gleason, M. E. J., Iida, M., Shrout, P. E., & Bolger, N. (2008). Receiving support as a mixed blessing: Evidence for dual effects of support on psychological outcomes. *Journal of Personality and Social Psychology*, 94, 824–838.
- Goodwin, M. S. (2012). Passive telemetric monitoring: Novel methods for real-world behavioral assessment. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 251–266). New York: Guilford Press.
- Gottman, J. M., & Roy, A. K. (1990). *Sequential analysis: A guide for behavioral researchers*. New York: Cambridge University Press.
- Green, A. S., Rafaeli, E., Bolger, N., Shrout, P. E., & Reis, H. T. (2006). Paper or plastic? Data equivalence in paper and electronic diaries. *Psychological Methods*, 11, 87–105.
- Gunthert, K. C. & Wenzel, S. J. (2012). Daily diary methods. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 144–159). New York: Guilford Press.
- Hays, R. B. (1989). The day-to-day functioning of close versus casual friendships. *Journal of Social and Personal Relationships*, 6, 21–37.

- Hedges, S. M., Jandorf, L., & Stone, A. A. (1985). Meaning of daily mood assessments. *Journal of Personality and Social Psychology*, 48, 428–434.
- Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life*. Thousand Oaks, CA: Sage Publications.
- Hinde, R. A. (1995). A suggested structure for a science of relationships. *Personal Relationships*, 2, 1–15.
- Hofmann, W., Baumeister, R. F., Förster, G., & Vohs, K. D. (2012). Everyday temptations: An experience sampling study of desire, conflict, and self-control. *Journal of Personality and Social Psychology*, 102, 1318–1335.
- Hormuth, S. E. (1986). The sampling of experiences in situ. *Journal of Personality*, 54, 262–293.
- Hormuth, S. E. (1990). *The ecology of the self: Relocation and self-concept change*. Cambridge: Cambridge University Press.
- Houts, A. C., Cook, T. D., & Shadish, W. R. (1986). The person-situation debate: A critical multiplist perspective. *Journal of Personality*, 54, 52–105.
- Ickes, W., & Tooke, W. (1988). The observational method: Studying the interaction of minds and bodies. In S. Duck & D. F. Hay (Eds.), *Handbook of personal relationships: Theory, research, and intervention* (pp. 79–97). Chichester: John Wiley & Sons.
- Intille, S. S. (2012). Emerging technology for studying daily life. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 267–282). New York: Guilford Press.
- Jensen-Campbell, L. A., & Graziano, W. G. (2000). Beyond the school yard: Relationships as moderators of daily interpersonal conflict. *Personality and Social Psychology Bulletin*, 26, 923–935.
- Kagan, J. (1984). *The nature of the child*. New York: Basic Books.
- Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, 4, 401–405.
- Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The Day

- Reconstruction Method. *Science*, 306, 1776–1780.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
- Kelley, H. H. (1997). The “stimulus field” for interpersonal phenomena: The source of language and thought about interpersonal events. *Personality and Social Psychology Review*, 1, 140–169.
- Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, 8, 115–128.
- Kernis, M. H., Cornell, D. P., Sun, C. R., Berry, A., & Harlow, T. (1993). There's more to self-esteem than whether it is high or low: The importance of stability of self-esteem. *Journal of Personality and Social Psychology*, 65, 1190–1204.
- Kessler, R. C. (1997). The effects of stressful life events on depression. *Annual Review of Psychology*, 48, 191–214.
- Knee, C. R., Canavello, A., Bush, A. L., & Cook, A. (2008). Relationship-contingent self-esteem and the ups and downs of romantic relationships. *Journal of Personality and Social Psychology*, 95, 608–627.
- Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, 99, 1042–1060.
- Lane, R. D., Zareba, W., Reis, H. T., Peterson, D. R., & Moss, A. J. (2011). Changes in ventricular repolarization duration during typical daily emotion in patients with Long QT Syndrome. *Psychosomatic Medicine*, 73, 98–105.
- Larsen, R. J. (1987). The stability of mood variability: A spectral analytic approach to daily mood assessments. *Journal of Personality and Social Psychology*, 52, 1195–1204.
- Larsen, R. J. (1990). Spectral analysis of psychological data. In V. Von Eye (Ed.), *Statistical methods in longitudinal research, Volume II: Time series and categorical longitudinal data* (pp. 319–349). Boston: Academic Press.
- Larsen, R. J., & Kasimatis, M. (1990). Individual-differences in entrainment of mood to the weekly calendar. *Journal of Personality and Social Psychology*, 58, 164–171.

- Larsen, R. J., & Kasimatis, M. (1991). Day-to-day physical symptoms: Individual-differences in the occurrence, duration, and emotional concomitants of minor daily illnesses. *Journal of Personality*, 59, 387–423.
- Larson, R. W., & Richards, M. H. (1994). *Divergent realities: The emotional lives of mothers, fathers, and adolescents*. New York: Basic Books.
- Larson, R. W., Richards, M. H., & Perry-Jenkins, M. (1994). Divergent worlds: The daily emotional experience of mothers and fathers in the domestic and public spheres. *Journal of Personality and Social Psychology*, 67, 1034–1046.
- Laurenceau, J. P., Barrett, L. F., & Rovine, M. J. (2005). The interpersonal process model of intimacy in marriage: A daily-diary and multilevel modeling approach. *Journal of Family Psychology*, 19, 314–323.
- Lawrence, D. M., & Schank, M. J. (1995). Health care diaries of young women. *Journal of Community Health Nursing*, 12, 171–182.
- Leary, M. R., Nezlek, J. B., Downs, D., Radford-Davenport, J., Martin, J., & McMullen, A. (1994). Self-presentation in everyday interactions: Effects of target familiarity and gender composition. *Journal of Personality and Social Psychology*, 67, 664–673.
- Leigh, B. C. (1993). Alcohol consumption and sexual activity as reported with a diary technique. *Journal of Abnormal Psychology*, 102, 490–493.
- Litt, M. D., Cooney, N. L., & Morse, P. (1998). Ecological momentary assessment (EMA) with treated alcoholics: Methodological problems and potential solutions. *Health Psychology*, 17, 48–52.
- Lucas, R. E. (2007). Adaptation and the set-point model of subjective well-being: Does happiness change after major life events? *Current Directions in Psychological Science*, 16, 75–79.
- Lucas, R. E., Le, K., & Dyrenforth, P. S. (2008). Explaining the extraversion/positive affect relation: Sociability cannot account for extraverts' greater happiness. *Journal of Personality*, 76, 385–414.
- Margolin, G., Christensen, A., & John, R. S. (1996). The continuance and spillover of everyday tension in distressed and nondistressed families. *Journal of Family Psychology*, 10, 304–321.
- McAdams, D. P. (1995). What do we know when we know a person? *Journal of*

Personality, 63, 365–396.

McClelland, D. C. (1957). Toward a science of personality psychology. In H. P. David & H. von Bracken (Eds.), *Perspective in personality theory* (355–382). New York: Basic Books.

McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review*, 96, 690–702.

McFarland, C., Ross, M., & Decourville, N. (1989). Women's theories of menstruation and biases in recall of menstrual symptoms. *Journal of Personality and Social Psychology*, 57, 522–531.

McGuire, W. J. (1997). Creative hypothesis generating in psychology: Some useful heuristics. *Annual Review of Psychology*, 48, 1–30.

Mehl, M. R., & Conner, T. S. (Eds.). (2012). *Handbook of research methods for studying daily life*. New York: Guilford Press.

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, 84, 857–870.

Mehl, M. R., & Robbins, M. L. (2012). The Electronically Activated Recorder (EAR). In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 176–192). New York: Guilford Press.

Mehl, M. R., Vazire, S., Ramirez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, 317, 82.

Menon, G. (1997). Are the parts better than the whole? The effects of decompositional questions on judgments of frequent behaviors. *Journal of Marketing Research*, 34, 335–346.

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, 7, 221–237.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102, 246–268.

Mohr, C. D., Armeli, S., Tennen, H., Carney, M. A., Affleck, G., & Hromi, A. (2001). Daily interpersonal experiences, context, and alcohol consumption: Crying in your beer and toasting good times. *Journal of Personality and*

Social Psychology, 80, 489–500.

- Moneta, G. B., & Csikszentmihalyi, M. (1996). The effect of perceived challenges and skills on the quality of subjective experience. *Journal of Personality*, 64, 275–310.
- Mortensen, C. R., & Cialdini, R. B. (2010). Full-cycle social psychology for theory and application. *Social and Personality Psychology Compass*, 4, 53–63.
- Moskowitz, D. S. (1994). Cross-situational generality and the interpersonal circumplex. *Journal of Personality and Social Psychology*, 66, 921–933.
- Moskowitz, D. S., & Sadikaj, G. (2012). Event-contingent recording. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 160–175). New York: Guilford Press.
- O'Connor, S. C., & Rosenblood, L. K. (1996). Affiliation motivation in everyday experience: A theoretical comparison. *Journal of Personality and Social Psychology*, 70, 513–522.
- Ong, A. D., Fuller-Rowell, T., & Burrow, A. L. (2009). Racial discrimination and the stress process. *Journal of Personality and Social Psychology*, 96, 1259–1271.
- Page-Gould, E., Mendoza-Denton, R., & Tropp, L. R. (2008). With a little help from my cross-group friend: Reducing anxiety in intergroup contexts through cross-group friendship. *Journal of Personality and Social Psychology*, 95, 1080–1094.
- Parkinson, B., Briner, R. B., Reynolds, S., & Totterdell, P. (1995). Time frames for mood: Relations between monetary and generalized ratings of affect. *Personality and Social Psychology Bulletin*, 21, 331–339.
- Pemberton, M. B., Insko, C. A., & Schopler, J. (1996). Memory for and experience of differential competitive behavior of individuals. *Journal of Personality and Social Psychology*, 71, 953–966.
- Pennebaker, J. W., & Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition & Emotion*, 10, 601–626.
- Penner, L. A., Shiffman, S., Paty, J. A., & Fritzsche, B. A. (1994). Individual differences in intraperson variability in mood. *Journal of Personality and*

Social Psychology, 66, 712–721.

- Pietromonaco, P. R., & Feldman-Barrett, L. (1997). Working models of attachment and daily social interactions. *Journal of Personality and Social Psychology*, 73, 1409–1423.
- Pinkus, R. T., Lockwood, P., Schimmack, U., & Fournier, M. A. (2008). For better and for worse: Everyday social comparison between romantic partners. *Journal of Personality and Social Psychology*, 95, 1180–1201.
- Pond, R. S., DeWall, C. N., Lambert, N. M., Deckman, T., Bonser, I. M., & Fincham, F. D. (2012). Repulsed by violence: Disgust sensitivity buffers trait, behavioral, and daily aggression. *Journal of Personality and Social Psychology*, 102, 175–188.
- Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15, 209–233.
- Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, 66, 3–8.
- Reis, H. T. (1994). Domains of experience: Investigating relationship processes from three perspectives. In R. Erber & R. Gilmour (Eds.), *Theoretical frameworks for personal relationships* (pp. 87–110). Hillsdale, NJ: Erlbaum.
- Reis, H. T. (2008). Reinvigorating the concept of situation in social psychology. *Personality and Social Psychology Review*, 12, 311–329.
- Reis, H. T., Collins, W. A., & Berscheid, E. (2000). The relationship context of human behavior and development. *Psychological Bulletin*, 126, 844–872.
- Reis, H. T., & Holmes, J. G. (2012). Perspectives on the situation. In K. Deaux & M. Snyder (Eds.), *The Oxford handbook of personality and social psychology* (pp. 64–92). New York: Oxford University Press.
- Reis, H. T., Lin, Y., Bennett, M. E., & Nezlek, J. B. (1993). Change and consistency in social participation during early adulthood. *Developmental Psychology*, 29, 633–645.
- Reis, H. T., Senchak, M., & Solomon, B. (1985). Sex differences in the intimacy of social interaction: Further examination of potential explanations. *Journal of*

Personality and Social Psychology, 48, 1204–1217.

- Reis, H. T., Sheldon, K. M., Gable, S. L., Roscoe, J., & Ryan, R. M. (2000). Daily well-being: The role of autonomy, competence, and relatedness. *Personality and Social Psychology Bulletin*, 26, 419–435.
- Reis, H. T., & Wheeler, L. (1991). *Studying social interaction with the Rochester Interaction Record*. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 24, pp. 269–318). San Diego, CA: Academic Press.
- Repetti, R. L. (1989). Effects of daily workload on subsequent behavior during marital interaction: The roles of social withdrawal and spouse support. *Journal of Personality and Social Psychology*, 57, 651–659.
- Ritter, J. M., & Langlois, J. H. (1988). The role of physical attractiveness in the observation of adult-child interactions: Eye of the beholder of behavioral reality? *Developmental Psychology*, 24, 254–263.
- Robinson, J. P., & Godbey, G. (1997). *Time for life: The surprising ways Americans use their time*. University Park: Pennsylvania State University Press.
- Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review*, 96, 341–357.
- Saxbe, D., & Repetti, R. L. (2010). For better or worse? Coregulation of couples' cortisol levels and mood states. *Journal of Personality and Social Psychology*, 98, 92–103.
- Sbarra, D. A. (2006). Predicting the onset of emotional recovery following nonmarital relationship dissolution: Survival analyses of sadness and anger. *Personality and Social Psychology Bulletin*, 32, 298–312.
- Schwartz, J. E., Neale, J., Marco, C., Shiffman, S. S., & Stone, A. A. (1999). Does trait coping exist? A momentary assessment approach to the evaluation of traits. *Journal of Personality and Social Psychology*, 77, 360–369.
- Schwarz, N. (2007). Retrospective and concurrent self-reports: The rationale for real-time data capture. In A. S. Stone, S. Shiffman, A. A. Atienza, & L. Nebeling (Eds.), *The science of real-time data capture* (pp. 11–26). New York: Oxford University Press.
- Schwarz, N. (2012). Why researchers should think “real-time”: A cognitive

- rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 22–42). New York: Guilford Press.
- Schwarz, N., Groves, R. M., & Schuman, H. (1998). Survey methods. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 1, 4th ed., pp. 143–179). New York: McGraw-Hill.
- Schwarz, N., & Sudman, S. (Eds.) (1996). *Answering questions: Methodology for determining cognitive and communicative processes in survey research*. San Francisco: Jossey-Bass.
- Shiffman, S., Paty, J. A., Gnys, M., Kassel, J. A., & Hickcox, M. (1996). First lapses to smoking: Within-subjects analysis of real-time reports. *Journal of Consulting and Clinical Psychology*, 64, 366–379.
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4, 1–32.
- Shulman, A. D., & Berman, H. J. (1975). Role expectations about subjects and experimenters in psychological research. *Journal of Personality and Social Psychology*, 32, 368–380.
- Skowronski, J. J., Betz, A. L., Thompson, C. P., & Shannon, L. (1991). Social memory in everyday life: Recall of self-events and other-events. *Journal of Personality and Social Psychology*, 60, 831–843.
- Sprecher, S. (1999). “I love you more today than yesterday”: Romantic partners’ perceptions of changes in love and related affect over time. *Journal of Personality and Social Psychology*, 76, 46–53.
- Stone, A. A., Hedges, S. M., Neale, J. M., & Satin, M. S. (1985). Prospective and cross-sectional mood reports offer no evidence of a blue Monday phenomenon. *Journal of Personality and Social Psychology*, 49, 129–134.
- Stone, A. A., Kessler, R. C., & Haythornthwaite, J. A. (1991). Measuring daily events and experiences: Decisions for the researcher. *Journal of Personality*, 59, 575–607.
- Stone, A. A., & Neale, J. M. (1984). Effects of severe daily events on mood. *Journal of Personality and Social Psychology*, 46, 137–144.
- Stone, A. A., Neale, J. M., & Shiffman, S. (1993). Daily assessments of stress and coping and their association with mood. *Annals of Behavioral Medicine*,

15, 8–16.

- Stone, A. A., Schwartz, J. E., Broderick, J. E., & Shiffman, S. S. (2005). Variability of momentary pain predicts recall of weekly pain: A consequence of the peak (or salience) memory heuristic. *Personality and Social Psychology Bulletin*, 31, 1340–1346.
- Stone, A. A., Schwartz, J. E., Neale, J. M., Shiffman, S., Marco, C. A. *et al.* (1998). A comparison of coping assessed by ecological momentary assessment and retrospective recall. *Journal of Personality and Social Psychology*, 74, 1670–1680.
- Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine*, 16, 199–202.
- Stone, A. A., Shiffman, S., Atienza, A. A., & Nebeling, L. (Eds.). (2007). *The science of real-time data capture*. New York: Oxford University Press.
- Stone, A. A., & Turkhan, J. S. (2000). Preface. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. ix–xi), Mahwah, NJ: Lawrence Erlbaum Associates.
- Suh, E., Diener, E., & Fujita, F. (1996). Only recent events matter. *Journal of Personality and Social Psychology*, 70, 1091–1102.
- Swim, J. K., Hyers, L. L., Cohen, L. L., Fitzgerald, D. C., & Bylsma, W. H. (2003). African American college students' experiences with everyday racism: Characteristics of and responses to these incidents. *Journal of Black Psychology*, 29, 38–67.
- Tennen, H., Suls, J., & Affleck, G. (1991). Personality and daily experience: The promise and the challenge. *Journal of Personality*, 59, 313–337.
- Thomas, D. L., & Diener, E. (1990). Memory accuracy in the recall of emotions. *Journal of Personality and Social Psychology*, 59, 291–297.
- Tidwell, M. C. O., Reis, H. T., & Shaver, P. R. (1996). Attachment, attractiveness, and social interaction: A diary study. *Journal of Personality and Social Psychology*, 71, 729–745.
- Tracy, J. L., Robins, R. W., & Sherman, J. W. (2009). The practice of psychological science: Searching for Cronbach's two streams in social-

- personality psychology. *Journal of Personality and Social Psychology*, 96, 1206–1225.
- Updegraff, J. A., Gable, S. L., & Taylor, S. E. (2004). What makes experiences satisfying? The interaction of approach-avoidance motivations and emotions in well-being. *Journal of Personality and Social Psychology*, 86, 496–504.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures*. Skokie, IL: Rand McNally.
- Wentland, E. J. (1993). *Survey responses: An evaluation of their validity*. New York: Academic Press.
- West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data-daily experiences. *Journal of Personality*, 59, 609–662.
- Wheeler, L., & Miyake, K. (1992). Social comparison in everyday life. *Journal of Personality and Social Psychology*, 62, 760–773.
- Wheeler, L., & Nezlek, J. B. (1977). Sex differences in social participation. *Journal of Personality and Social Psychology*, 35, 742–754.
- Wheeler, L., & Reis, H. T. (1991). Self-recording of everyday life events: Origins, types, and uses. *Journal of Personality*, 59, 339–354.
- Wilhelm, F. H., Grossman, P., & Müller, M. I. (2012). Bridging the gap between the laboratory and the real world: Integrative ambulatory psychophysiology. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 210–234). New York: Guilford Press.
- Williams, K. J., Suls, J., Alliger, G. M., Learner, S. M., & Wan C. K. (1991). Multiple role juggling and daily mood states in working mothers: An experience sampling study. *Journal of Applied Psychology*, 76, 664–674.
- Woike, B. A. (1995). Most-memorable experiences: Evidence for a link between implicit and explicit motives and social cognitive processes in everyday life. *Journal of Personality and Social Psychology*, 68, 1081–1091.
- Wong, M. M., & Csikszentmihalyi, M. (1991). Affiliation motivation and daily experience: Some issues on gender differences. *Journal of Personality and Social Psychology*, 60, 154–164.
- Wood, J. V., Saltzberg, J. A., Neale, J. M., Stone, A. A., & Rachmiel, T. B. (1990). Self-focused attention, coping responses, and distressed mood in

everyday life. *Journal of Personality and Social Psychology*, 58, 1027–1036.

Wortman, C. B., & Silver, R. D. (1989). The myths of coping with loss. *Journal of Consulting and Clinical Psychology*, 57, 349–357.

Yip, T. (2005). Sources of situational variation in ethnic identity and psychological well-being: A palm pilot study of Chinese American students. *Personality and Social Psychology Bulletin*, 31, 1603–1616.

¹ Here and elsewhere we refer to retrospections about events in one's relatively recent, adult experience. The limitations of retrospection for recalling early life experience are even greater (Brewin, Andrews, & Gotlib, 1993).

² In this instance, event-contingent methods may be preferable to signal-contingent methods, which usually allow participants to turn off the signaling device when intrusion would be unacceptable. Although this provision is needed to obtain cooperation, event records, which can be provided after the fact, would not omit the event.

³ Disaggregation, or treating each record (i.e., day, event, signal) as an independent entry, is not recommended. This in essence ignores nesting in the data structure. Aside from the prospect of introducing substantial bias by ignoring dependency, this approach bypasses the most interesting feature of daily experience research – variation across different individuals.

⁴ This is because heteroscedasticity engenders systematic discrepancies in the degree to which each person's observed mean deviates from his or her true mean.

⁵ This interpretation of b_0 is contingent on centering each person's stress values around their mean, which is commonly desirable (Bryk & Raudenbush, 1992). Thus, in the example that follows, day-level predictors (stress) would be centered on the person's mean level of stress across the study; person-level predictors (self-esteem) would be centered on the sample mean. The resulting values are therefore interpretable as deviations from the corresponding mean.

⁶ We note that this analysis assumes independence of records at the lowest level of analysis. If not (as when husbands and wives are describing the same events), it is necessary to take this dependence into account. Bolger and Laurenceau (in press) discuss these methods in detail.

⁷ Adding more subjects can increase power at level 2, but power at level 1 also depends on the number of within-subject assessments, so, for example, having a huge sample with only two within-person observations will not provide enough power to test within-person processes.

⁸ It should be noted that sequential analysis is facilitated by equally, or at least systematically, spaced intervals. The irregular gaps of signal-contingent and especially event-contingent sampling may complicate computation (by requiring specification of the length of each interval) and interpretation of lags and sequences, unless data are aggregated over fixed intervals prior to analysis.

Chapter sixteen Survey Research

Jon A. Krosnick, Paul J. Lavrakas and Nuri Kim

When social psychologists see a chapter offering to tell them how to conduct survey research, some respond by saying: “I don't do surveys, so the survey research methodology literature doesn't offer me tools of value for my research program. I do experiments, because they offer me the opportunity to document causality definitively. Surveys provide merely correlational data with no real value for inferring causal influence, so that's not my thing.”

Such a statement reveals a lack of understanding of what survey research methodology involves. Surveys are not inherently correlational. Instead, surveys are defined as data collections for which a researcher (1) defines a specific population of people to be described, (2) draws a systematic and representative sample of members of the population, (3) collects data from those individuals either by asking them questions or by asking them to perform other tasks, and (4) computes statistics that properly reflect the nature of the sampling process used to select the individuals.¹ In fact, experiments can and routinely are embedded in surveys. Thus, surveys are not antithetical to experiments or “merely correlational.” Instead, the huge literature on survey research methods offers social psychologists the opportunity to do much of the research they wish to do even more effectively, because that literature offers insights that will improve the value of the field's experimental and nonexperimental studies. More than ever before, social psychologists stand to benefit from having a command of the survey research literature and of the technique of survey research, for a variety of specific reasons.

First, a great deal of research in social psychology is done with questionnaires, and the design of those questionnaires is typically done either by reusing a questionnaire that another investigator used previously or by designing a new questionnaire based on intuition. Yet a large and growing literature in survey research suggests how to design questionnaires so as to yield measurements with maximum reliability and validity. Therefore, social psychologists' assessments can become more precise, and, by minimizing the distorting impact of random and systematic measurement error, researchers can maximize their chances of

detecting real effects and associations with minimal sample sizes.

Second, if social psychology's main goal is to produce findings that will engage the interest of college undergraduates enrolled in the courses that we teach, then continuing to collect the vast majority of our data from such students is wise. That is, if we seek mainly to fill textbooks to sell to undergraduates as we teach them about themselves, basing our studies on people like them will make the findings especially compelling. But social psychologists are increasingly called upon to provide expert testimony in court, to advise government agencies, to consult with corporations (about how to manage their workforces and how to design and sell their products and services), to advise political candidates (about how to win elections), to consult with political interest groups (about how to influence government policy making), and to offer insights on human behavior via the news media. In such settings, the findings of our research are sometimes scrutinized by people and organizations who wish to dispute our conclusions on nonscientific grounds. Consequently, it is especially important that the research methods and resulting data on which our opinions are based provide a convincing justification for generalizing our findings beyond the subpopulation of college students.

Equally important, many social psychologists seek to have our research funded by government agencies, such as the National Science Foundation and the National Institutes of Health, and by private foundations. To justify investment of substantial funds in our work, it may be important that the work provide an empirical justification for making claims about a wide range of persons at all stages in the life cycle and living in many different social settings. Presuming that our findings generalize from college students to the broader population in the absence of supportive data is unlikely to be convincing to people making tough decisions about resource allocation. Therefore, we owe it to ourselves to confirm the generalizability of our findings by at least sometimes collecting data from broader segments of the public.

One increasingly popular approach to doing so is to post an experiment on a website and to allow any interested visitor to provide data (Chapter 17 by Maniaci and Rogge in this volume discusses methods for conducting research on the Internet). This may seem attractive, because the demographics of the participants are often obviously more diverse than college students. But a researcher carrying out such a study does not know the identities of the particular people who chose to do the study, how they happened to learn about its availability on the Web, whether the same person participated in the study

multiple times, and the degree to which the entire group of participating respondents match a specifiable population (e.g., all Americans).

Another popular approach is to hire respondents through services such as Amazon's Mechanical Turk. These individuals complete an experiment in exchange for a small amount of money (Maniaci & Rogge, Chapter 17 in this volume). Yet it is again not clear who these people are and the degree to which they match a specific population of adults. Most notably, because of the nature of sources like Mechanical Turk, the participants seem unlikely to include people whose lives do not encourage them to earn modest sums by performing small tasks online. Indeed, some online experiments might be completed quite often by college students, so that the apparent diversification of the respondent pool is more of an illusion than a reality.

Put simply, using such techniques for data collection transforms social psychologists from knowing that their samples of college student participants are not representative of the larger population to not knowing the extent to which their participants are representative and having no scientific basis for making generalizations. In other words, a researcher would continue to be in a position where he or she must ask outsiders to “trust me” – to grant that his or her findings are broadly generalizable, despite the absence of strong evidence to that effect.

For all of these reasons, social psychologists have an incentive to do some of our studies with representative samples of the populations we seek to describe, so that we can have a scientific basis for making claims wherein we generalize our findings to populations. This is not to say that studies of college students are without merit. They are valuable, and they should continue to be done, as should studies of haphazard samples of other types of participants, because such work yields valuable scientific insights. But that work will be even more persuasive, insightful, and constructive if social psychologists occasionally replicate and extend their (experimental and nonexperimental) studies by collecting data from representative samples of defined populations.

This recommendation is very much in line with the observation offered by Petty and Cacioppo (1996), who noted that a laboratory study of college students “examines the viability of some more general hypothesis about the relationship between two (or more) variables and ascertains what might be responsible for this relationship. Once the relationship is validated in the laboratory, its applicability to various specific situations and populations can be ascertained” (pp. 3–4). In order to do so with regard to any specific population (even the

population of American college students), social psychologists must understand and employ the techniques of sampling that are the bailiwick of survey research.

Fortunately, carrying out (experimental and nonexperimental) survey research to supplement laboratory-based studies is probably much easier than many social psychologists realize. First, TESS (Time-Sharing Experiments for the Social Sciences, www.tessexperiments.org) is a platform funded by the National Science Foundation that allows any researcher to conduct experiments embedded in surveys of representative samples of American adults at no cost. Similar opportunities are offered by other organizations in other countries (e.g., <http://www.centerdata.nl/en/>).

Second, many survey data sets are made available to researchers at no cost, which include measures of interest to social psychologists, including the General Social Survey, the American National Election Studies, and many other surveys available through archives such as the Interuniversity Consortium for Political and Social Research (<http://www.icpsr.umich.edu>) and the Roper Center for Public Opinion Research (<http://www.ropercenter.uconn.edu/>). Analyzing data from secondary sources is increasingly of interest to psychologists (Trzesniewski, Donnellan, & Lucas, 2010). To properly analyze such data, social psychologists must understand how they were collected and must therefore understand the basics of survey methodology.

Third, funding agencies are increasingly willing to pay the costs of primary data collection from representative samples by social psychologists. Obtaining such funding is likely to improve the *apparent* scientific value of a proposed line of work in the eyes of non-psychologists, by permitting the empirical documentation of the generalizability of findings to populations of interest, therefore enhancing the fundability of a line of investigation, rather than decreasing it. That is, by conducting some experiments with samples that are representative of a population, social psychologists can reassure skeptics that their findings do indeed generalize. To effectively propose to conduct such work, a social psychologist can recruit a survey expert to join his or her research team, but the social psychologist can also choose to learn the insights offered by the survey methodology literature. This chapter is designed for both sorts of scholars.

Specifically, the survey research literature offers guidance to social psychologists with regard to: (1) how to collect data optimally in one of four modes (face-to-face interviews, telephone interviews, paper-and-pencil questionnaires, and electronic questionnaires delivered via a computer); (2) how

to draw a sample of respondents from the population for data collection; (3) how to hire, train, and supervise interviewers (when data are collected via face-to-face or telephone interviews); (4) how to design and pretest questionnaires; (5) how to manage the data collection process; and (6) how to conduct proper statistical analyses to permit generalization of findings in light of the particular sampling approach used in a study. Whether a social psychologist wishes to manage the conduct of his or her survey fully or chooses instead to hire a firm to collect the data, full understanding of the particulars of the survey method is essential. Therefore, this chapter is designed to help social psychologists understand the logic of survey research and to benefit from the insights that professionals in the field have gained about best practices.

Defined formally, *survey research* is a specific type of field study that involves the collection of data from a sample of elements drawn systematically to be representative of a well-defined, large, and geographically diverse population (e.g., all adult women living in the United States) often, though not necessarily, through the use of a questionnaire (for more lengthy discussions, see Babbie, 1990; Fowler, 2009; Frey, 1989; Lavrakas, 1993; Sapsford, 2007; Weisberg, Krosnick, & Bowen, 1996). In order to understand how to conduct such research more effectively and efficiently to accurately describe people's thinking and action, survey researchers have done extensive research on various aspects of survey methodology. This chapter reviews high points of that literature, outlining more specifically why survey research may be valuable to social psychologists, explaining the utility of various study designs, reviewing the basics of survey sampling and questionnaire design, and describing optimal procedures for data collection.

Study Designs

A survey-based research project can employ a variety of different designs, each of which is suitable for testing hypotheses of interest to social psychologists. In this section we review several standard designs, including cross-sectional, repeated cross-sectional, panel, and mixed designs, and discuss when each is appropriate for social psychological investigation.

Cross-Sectional Surveys

Cross-sectional surveys involve the collection of data at a single point in time from a sample drawn from a specified population. This design can be used by

social psychologists for documenting the prevalence of particular characteristics in a population. For example, researchers studying altruism might want to begin an investigation by documenting the frequency with which people report altruistic behaviors. Or researchers studying aggression might wish to begin their work by documenting the frequency with which people report aggressive behaviors, in order to provide a compelling initial backdrop for their in-depth study of aggressiveness.

Cross-sectional surveys can also yield correlational evidence about the directions and magnitudes of associations between pairs of variables. Such correlations do not themselves provide evidence of the causal processes that gave rise to them. But such correlations are informative about the plausibility of a causal hypothesis. That is, if variable A is thought to be a cause of variable B but the two turn out to be uncorrelated empirically, the plausibility of the causal claim is thus diminished.

Cross-sectional surveys also offer the opportunity to test causal hypotheses in a number of ways. For example, using statistical techniques such as two-stage least squares regression, it is possible to estimate the causal impact of variable A on variable B, as well as the effect of variable B on variable A (Blalock, 1972). Such an analysis rests on important assumptions about causal relations among variables, and these assumptions can be tested and revised as necessary (see, e.g., James & Singh, 1978). Furthermore, path analytic techniques can be applied to test hypotheses about the mediators of causal relations (Baron & Kenny, 1986; Kenny, 1979), thereby validating or challenging notions of the psychological mechanisms involved. And cross-sectional data can be used to identify the moderators of relations between variables, thereby also shedding some light on the causal processes at work (e.g., Krosnick, 1988b). (For discussions of both mediators and moderators, see Judd, Yzerbyt, & Muller, Chapter 25 in this volume.) For example, consider the hypothesis that a perceiver will evaluate another person in part based on the degree to which that person holds similar attitudes. That is, attitude similarity is thought to cause attraction. An initial test of this hypothesis might be afforded by gauging whether perceivers who are more in favor of strict gun control laws are more attracted to a political candidate who also favors strict gun control laws. This can be gauged by the cross-sectional correlation between perceivers' attitudes on gun control and liking of the candidate.

But such a correlation could be attributable to the influence of attitude similarity on attraction or to the influence of attraction to the candidate on

attitude similarity. That is, a perceiver might like a candidate because they share the same political party affiliation, and the candidate's articulate endorsement of strict gun control attitudes might then convince the perceiver to adjust his or her own attitude on the issue to match the candidate's. If this latter process were to be true, we would expect it to be more common among people with weak attitudes toward gun control laws, whereas deriving liking of the candidate from similarity of attitudes on gun control would presumably be more common among people whose gun control attitudes are strong (see Krosnick, 1988a). Therefore, by exploring whether the strength of the association between gun control attitude similarity and candidate liking varies with the strength of the perceiver's gun control attitude, we can generate evidence consistent with one or the other or neither of these causal claims (Krosnick, 1988b).

A single, cross-sectional survey can also be used to assess the impact of a social event. For example, Krosnick and Kinder (1990) studied priming in a real-world setting by focusing on the Iran-Contra scandal. On November 25, 1986, the American public learned that members of the National Security Council had been funneling funds (earned through arms sales to Iran) to the Contras fighting to overthrow the Sandinista government in Nicaragua. Although there had been almost no national news media attention to Nicaragua and the Contras previously, this revelation led to a dramatic increase in the salience of that country in the American press during the following weeks. Krosnick and Kinder suspected that this coverage might have primed Americans' attitudes toward U.S. involvement in Nicaragua and thereby increased the impact of these attitudes on evaluations of President Ronald Reagan's job performance.

To test this hypothesis, Krosnick and Kinder (1990) took advantage of the fact that data collection for the 1986 National Election Study, a national survey, was underway well before November 25 and continued well after that date. So these investigators simply split the survey sample into one group of respondents who had been interviewed before November 25 and another group consisting of those who had been interviewed afterward. As expected, overall assessments of presidential job performance were based much more strongly on attitudes toward U.S. involvement in Nicaragua in the second group than they were in the first group. This use of survey data amounts to employing them to "create" a quasiexperiment.

Furthermore, Krosnick and Kinder (1990) found that this priming effect was concentrated primarily among people who were not especially knowledgeable about politics (so-called political novices), a finding permitted by the

heterogeneity in political expertise in a national sample of adults. From a psychological viewpoint, this suggests that news media priming occurs most when opinions and opinion formation processes are not firmly grounded in past experience and in supporting knowledge bases. From a political viewpoint, this finding suggests that news media priming may not be especially politically consequential in nations where political expertise is high throughout the population.

Repeated Cross-Sectional Surveys

One drawback of the study by Krosnick and Kinder (1990) is that splitting a national survey sample in two parts confounds time with sample attributes. That is, the sample of respondents interviewed before November 25 is likely to have had some characteristics that distinguish them from those interviewed after November 25. Specifically, the former individuals may have been home more often and/or may have been more willing to agree to be interviewed. This leaves open the possibility that the two groups of people differed from one another not only because of the revelation of the Iran-Contra affair but for other reasons as well. This confounding can be overcome by conducting multiple independent surveys, one before an event occurs and one after. That way, the survey samples will be comparable to one another, so the impact of time can be studied more clearly.

Furthermore, the conduct of multiple independent surveys (drawing representative samples from the same population) over time offer the opportunity to generate a different type of evidence consistent with a hypothesized causal relation by assessing whether changes over time in a dependent variable parallel changes in a proposed independent variable. If a hypothesized causal relation exists between two variables, between-wave changes in the independent variable should be mirrored by between-wave changes in the dependent variable. For example, if one believes that interracial contact may reduce interracial prejudice, an increase in interracial contact over a period of years in a society should be paralleled by or should precede a reduction in interracial prejudice.

One study along these lines was reported by Schuman, Steeh, and Bobo (1985). Using cross-sectional surveys conducted between the 1940s and the 1980s in the United States, these investigators documented dramatic increases in the prevalence of positive attitudes toward principles of equal treatment of whites and blacks. And there was every reason to believe that these general

principles might be important determinants of people's attitudes toward specific government efforts to ensure equality. However, there was almost no shift during these years in public attitudes toward specific implementation strategies. This challenges the notion that the latter attitudes were shaped powerfully by the general principles of equal treatment of whites and blacks.

Repeated cross-sectional surveys can also be used to study the impact of social events that occurred between the surveys (e.g., Kam & Ramos, 2008; Weisberg, Haynes, & Krosnick, 1995). And repeated cross-sectional surveys can be combined into a single data set for statistical analysis, using information from one survey to estimate parameters in another survey (e.g., Brehm & Rahn, 1997; Kellstedt, Peterson, & Ramirez, 2010).

Panel Surveys

In a panel survey, data are collected from the same people at two or more points in time. One use of panel data is to assess the stability of psychological constructs and to identify the determinants of stability (e.g., Krosnick, 1988a; Krosnick & Alwin, 1989; Trzesniewski, Donnellan, & Robins, 2003). Just as with a single survey, one can gauge the prevalence of a characteristic in the population and cross-sectional associations between variables. But one can also test causal hypotheses in at least two ways. First, a researcher can examine whether individual-level changes over time in an independent variable correspond to individual-level changes in a dependent variable over the same period of time. So, for example, one can ask whether people who experienced increasing interracial contact manifested decreasing racial prejudice, while at the same time people who experienced decreasing interracial contact manifested increasing racial prejudice.

Second, one can assess whether changes over time in a dependent variable can be predicted by prior levels of an independent variable. So, for example, do people who had the highest amounts of interracial contact at Time 1 manifest the largest decreases in racial prejudice between Time 1 and Time 2? Such a demonstration provides relatively strong evidence consistent with a causal hypothesis, because the changes in the dependent variable could not have caused the prior levels of the independent variable (e.g., Blalock, 1985; Kessler & Greenberg, 1981 on the methods; see Chanley, Rudolph, & Rahn, 2000; Eveland, Hayes, Shah, & Kwak, 2005; Rahn, Krosnick, & Breuning, 1994 for an illustration of its application).

One application of this approach occurred in a study of a long-standing social

psychological idea called the *projection hypothesis*. Rooted in cognitive consistency theories, it proposes that people may overestimate the extent to which they agree with others whom they like, and they may overestimate the extent to which they disagree with others whom they dislike. By the late 1980s, a number of cross-sectional studies by political psychologists yielded correlations consistent with the notion that people's perceptions of the policy stands of presidential candidates were distorted to be consistent with attitudes toward the candidates (e.g., Granberg, 1985; Kinder, 1978) .

However, there were alternative theoretical interpretations of these correlations, so an analysis using panel survey data seemed in order. Krosnick (1991a) did just such an analysis exploring whether attitudes toward candidates measured at one time point could predict subsequent shifts in perceptions of presidential candidates' issue stands. He found no projection at all to have occurred, thereby suggesting that the previously documented correlations were more likely attributable to other processes (e.g., deciding how much to like a candidate based on agreement with him or her on policy issues; Byrne, 1971; Krosnick, 1988b).

The impact of social events can be gauged especially powerfully with panel data. For example, Krosnick and Brannon (1993) studied news media priming using such data. Their interest was in the impact of the Gulf War on the ingredients of public evaluations of presidential job performance. For the 1990–1991 National Election Panel Study of the Political Consequences of War, a panel of respondents had been interviewed first in late 1990 (before the Gulf War) and then again in mid-1991 (after the war). The war brought with it tremendous news coverage of events in the Gulf, and Krosnick and Brannon suspected that this coverage might have primed attitudes toward the Gulf War, thereby increasing their impact on public evaluations of President George H. W. Bush's job performance. This hypothesis was confirmed by comparing the determinants of presidential evaluations in 1990 and 1991. Because the same people had been interviewed on both occasions, this demonstration is not vulnerable to a possible alternative explanation of the Krosnick and Kinder (1990) results described earlier: that different sorts of people were interviewed before and after the Iran-Contra revelation.

Panel surveys do have some disadvantages. First, although people are often quite willing to participate in a single cross-sectional survey, fewer are usually willing to complete multiple interviews. Furthermore, with each additional wave of panel data collection, it becomes increasingly difficult to locate respondents to

reinterview them, because some people move to different locations, some die, and so on. This attrition may threaten the representativeness of panel survey samples if the members of the first-wave sample who agree to participate in several waves of data collection differ in meaningful ways from the people who are interviewed initially but do not agree to participate in subsequent waves of interviewing. However, studies of panel attrition have generally found little impact of attrition on sample representativeness and substantive results (Alderman et al., 2001; Beckett, Gould, Lillard, & Welch, 1988; Clinton, 2001; Falaris & Peters, 1998; Fitzgerald, Gottschalk, & Moffitt, 1998a, 1998b; Watson, 2003; Zabel, 1998; Zagorsky & Rhoton, 1999; Ziliak & Kniesner, 1998).

Also, participation in the initial survey may sensitize respondents to the issues under investigation, thus changing the phenomena being studied. As a result, respondents may give special attention or thought to these issues, which may have an impact on subsequent survey responses. For example, Bridge *et al.* (1977) demonstrated that individuals who participated in a survey interview about health subsequently considered the topic to be more important. And this increased importance of the topic can be translated into changed behavior. For example, people interviewed about politics are subsequently more likely to vote in elections (Granberg & Holmberg, 1992; Kraut & McConahay, 1973; Voogt & Van Kempen, 2002; Yalch, 1976). Even answering a single survey question about one's intention to vote can increase the likelihood that an individual will turn out to vote on election day (Greenwald, Carnot, Beach, & Young, 1987; cf. Mann, 2005).

Finally, panel survey respondents may want to appear consistent in their responses across waves. Therefore, people may be reluctant to report opinions or behaviors that appear inconsistent with what they recall having reported during earlier interviews. The desire to appear consistent could mask genuine changes over time.

Combined Use of Cross-Sectional and Panel Surveys

Researchers can capitalize on the strengths of each of the aforementioned designs by incorporating both cross-sectional and panel surveys into a single study. If, for example, a researcher is interested in conducting a two-wave panel survey but is concerned about carryover effects, he or she could conduct an additional cross-sectional survey at the second wave. That is, the identical questionnaire could be administered to both the panel respondents and to an

independent sample drawn from the same population. Significant differences between the data collected from these two samples would suggest that carryover effects were, in fact, a problem in the panel survey. In effect, the cross-sectional survey respondents can serve as a “control group” against which panel survey respondents can be compared.

Experiments within Surveys

Additional evidence of causal processes can be documented in surveys by building in experiments. If respondents are randomly assigned to “treatment” and “control” groups, differences between the two groups can then be attributed to the treatment. Every one of the survey designs described in the preceding sections can be modified to incorporate experimental manipulations. Some survey respondents (assigned randomly) can be exposed to one version of a questionnaire, whereas other respondents are exposed to another version. Differences in responses can then be attributed to the specific elements that were varied.

Many of the elements of traditional laboratory experiments can be easily implemented in the context of surveys, especially online surveys (see Maniaci & Rogge, Chapter 17 in this volume). For example, many social psychological studies exposed participants to a persuasive message (either in print, orally, or via video), and then participants answered questions measuring dependent variables. Such persuasive messages can easily be presented to respondents completing online surveys. Furthermore, telephone interviewers can read the persuasive message aloud to their respondents, and face-to-face interviewers can do the same, or can present a print message on a piece of paper, or can use their laptops to display a video presentation of the message.²

It is also possible to set up online networks of respondents who interact with one another in a group discussion in the context of an online survey. And such a group discussion can also be implemented with fictitious other respondents whose “behavior” is controlled by a computer (Davies & Gangadharan, 2009). Thus, group interactions that might be implemented in the lab can also be implemented with a representative sample of respondents.

Many social psychologists are aware of examples of survey research that have incorporated experiments to explore effects of question order and question wording (see, e.g., Box-Steffensmeier, Jacobson, & Grant, 2000; Couper, Traugott, & Lamias, 2001; Schuman & Presser, 1981). Less salient are the

abundant examples of experiments within surveys that have been conducted to explore other social psychological phenomena.

Racism.

One experimental study within a survey was reported by Kinder and Sanders (1990), who were interested in the impact of public debates on public opinion on affirmative action. Sometimes, opponents of affirmative action have characterized it as entailing reverse discrimination against qualified white candidates; other times, opponents have characterized affirmative action as giving unfair advantages to minority candidates. Did this difference in framing change the way the general public formed opinions on the issue?

To answer this question, Kinder and Sanders (1990) asked white respondents in a national survey about whether they favored or opposed affirmative action programs in hiring and promotions and in college admissions. Some respondents were randomly assigned to receive a description of opposition to affirmative action as emanating from the belief that it involves reverse discrimination. Other respondents, again assigned randomly, were told that opposition to affirmative action emanates from the belief that it provides unfair advantages to minorities.

This experimental manipulation of the framing of opposition did not alter the percentages of people who said they favored or opposed affirmative action, but it did alter the processes by which those opinions were formed. When affirmative action was framed as giving unfair advantage to minorities (thereby making minority group members salient), it evoked more anger, disgust, and fury from respondents, and opinions were based more on general racial prejudice, on intolerance of diversity in society, and on belief in general moral decay in society. But when affirmative action was framed as reverse discrimination against qualified whites (thereby making whites more salient), opinions were based more on the perceived material interests of the respondent and of whites as a group.

Because Kinder and Sanders (1990) analyzed data from a national survey, respondents varied a great deal in terms of their political expertise. Capitalizing on this diversity, Kinder and Sanders found that the impact of framing was concentrated nearly exclusively among political novices. This reinforced the implication of Krosnick and Kinder's (1990) finding regarding political expertise in their research on news media priming described earlier.

Sniderman and Tetlock (1986) and Sniderman, Tetlock and Peterson (1993)

have also conducted experiments within surveys to assess whether conservative values encourage racial prejudice in judgments about who is entitled to public assistance and who is not. In their studies, respondents were told about a hypothetical person in need of public assistance. Different respondents were randomly assigned to receive different descriptions of the person, varying in terms of previous work history, marital and parental status, age, and race. Interestingly, conservatives did not exhibit prejudice against blacks when deciding whether he or she should receive public assistance, even when the person was said to have violated traditional values (e.g., by being a single parent or having a history of being an unreliable worker). In fact, when presented with an individual who had a history of being a reliable worker, conservatives were substantially more supportive of public assistance for blacks than for whites. However, conservatives were significantly less supportive of public policies designed to assist blacks as a group and were more likely to believe that blacks are irresponsible and lazy. Sniderman and Tetlock (1986) concluded that a key condition for the expression of racial discrimination is therefore a focus on groups rather than individual members of the groups, and that a generally conservative orientation does not encourage individual-level discrimination .

Mood and Life Satisfaction.

Schwarz and Clore (1983) conducted an experiment in a survey to explore mood and misattribution. They hypothesized that general affective states can sometimes influence judgments via misattribution. Specifically, these investigators presumed that weather conditions (sunny vs. cloudy) influence people's moods, which in turn may influence how happy they say they are with their lives. This presumably occurs because people misattribute their current mood to the general conditions of their lives rather than to the weather conditions that happen to be occurring when they are asked to make the judgment. As a result, when people are in good moods, they may overstate their happiness with their lives.

To test this hypothesis, Schwarz and Clore (1983) conducted telephone interviews with people on either sunny or cloudy days. Among respondents who were randomly assigned to be asked simply how happy they were with their lives, those interviewed on sunny days reported higher satisfaction than people interviewed on cloudy days. But among people randomly assigned to be asked first, "By the way, how's the weather down there?", those interviewed on sunny days reported identical levels of life satisfaction to those interviewed on cloudy

days. The question about the weather presumably led people to properly attribute some of their current mood to current weather conditions, thereby insulating subsequent life satisfaction judgments from influence.

The Benefits of Experiments within Surveys.

What is the benefit of doing these experiments in representative sample surveys? Couldn't they instead have been done just as well in laboratory settings with college undergraduates? Certainly, the answer to this latter question is yes; they could have been done as traditional social psychological experiments. But the value of doing the studies within representative sample surveys is at least threefold. First, survey evidence documents that the phenomena are widespread enough to be observable in the general population. This bolsters the apparent value of the findings in the eyes of the many non-psychologists who instinctively question the generalizability of laboratory findings regarding undergraduates.

Second, estimates of effect sizes from surveys provide more accurate bases for assessing the significance that any social psychological process is likely to have in the course of daily life. Effects that seem large in the lab (perhaps because undergraduates are easily influenced) may actually be quite small and thereby much less socially consequential in the general population.

Third, general population samples allow researchers to explore whether attributes of people that are homogeneous in the lab but vary dramatically in the general population (e.g., age, educational attainment) moderate the magnitudes of effects or the processes producing them (e.g., Kinder & Sanders, 1990).

Implicit Measurement

Social psychologists are increasingly interested in implicit measurement and are able to implement implicit assessment procedures easily in laboratory settings (see Gawronski & De Houwer, Chapter 12 in this volume, for an introduction to implicit methods). Fortunately, many such procedures can be implemented in the context of surveys as well. For example, it is now routine for computers used by face-to-face interviewers and telephone interviewers and used by respondents for Internet surveys to record the amount of time each respondent takes to answer each question. Such data have proven to be quite valuable analytically. Furthermore, procedures such as the Implicit Association Test and the Affect Misattribution Paradigm are computer-based and can therefore easily be incorporated in survey data collection done via the Internet or via laptop

computers that face-to-face interviewers bring to respondents' homes (e.g., Pasek et al., 2009; Payne et al., 2010).

For example, Pasek and colleagues (2009) and Payne and colleagues (2010) implemented the Affect Misattribution Paradigm within the context of online surveys of representative national samples of American adults. These studies assessed anti-black attitudes and compared implicit assessments with explicit assessments using traditional measures of constructs, such as stereotypes and symbolic racism. Statistical analyses documented negative associations of implicit and explicit anti-black attitudes with voting for Barack Obama in the 2008 U.S. presidential election and positive associations with voting for John McCain. The impact of implicit attitudes was partly but not completely mediated by explicit attitudes. The same findings were obtained in analyses of data collected via face-to-face interviews with a representative national sample of American adults in their homes in 2008 .

Sampling

Once a survey design has been specified, the next step is selecting a sampling method (see, e.g., Henry, 1990; Kalton, 1983; Kish, 1965; Sudman, 1976). The social science literature describes many examples where the conclusions of studies were dramatically altered when proper sampling methods were used (see, e.g., Laumann, Michael, Gagnon, & Michaels, 1994). In this section we explain a number of sampling methods and discuss their strengths and weaknesses. In this discussion the term “element” is used to refer to the individual unit about which information is sought. In most studies, elements are the people who make up the population of interest, but elements can also be groups of people, such as families, corporations, or departments.³ A *population* is the complete group of elements to which one wishes to generalize findings obtained from a sample.

Probability Sampling

There are two general classes of sampling methods: nonprobability and probability sampling. *Nonprobability sampling* refers to selection procedures in which elements are not randomly selected from the population or some elements have unknown probabilities of being selected. *Probability sampling* refers to selection procedures in which elements are randomly selected from the sampling frame (usually the population of interest), and each element has a known, nonzero chance of being selected. This does not require that all elements have an

equal probability, nor does it preclude some elements from having a certain (1.00) probability of selection. However, it does require that the selection of each element must be independent of the selection of every other element.

Probability sampling affords two important advantages. First, researchers can be confident that a selected sample is representative of the larger population from which it was drawn only when a probability sampling method has been used. When elements have been selected through other procedures or when portions of the population had no chance of being included in the sample, there is no way to know whether the sample is representative of the population. Generalizations beyond the specific elements in the sample are therefore only warranted when probability sampling methods have been used.

The second advantage of probability sampling is that it permits researchers to precisely estimate the amount of variance present in a given dataset that is attributable to sampling error. That is, researchers can calculate the degree to which random differences between the sample and the sampling frame are likely to have diminished the precision of the obtained estimates. Probability sampling also permits researchers to construct confidence intervals around their parameter estimates, which indicate the precision of the point estimates.

It might seem as if social psychologists need not understand how sampling processes work – perhaps they can just rely on survey professionals to design the sample and collect the data for them, and the psychologists can simply analyze the data while trusting its representativeness. But in fact, this is not true. Many probability sampling designs incorporate complexities that cause unequal probabilities of selection and clustering. These unequal probabilities and clustering must be taken into account when doing statistical analysis in order to avoid introducing bias. Furthermore, in order to assure that a representative sample accurately describes the population of interest, survey professionals routinely compute post-stratification weights to enhance the match of the sample to the population. Such weights are needed because even when a random sample is drawn from a population, the sample of people who complete the survey often deviates in observable ways from the population. For example, national surveys of random samples of Americans routinely overrepresent well-educated people and women, because members of these groups are more willing to participate than are less educated people and men. Because the true distributions of education and sex in the population can be known from the U.S. Census, post-stratification weights can be applied to a set of survey data to adjust the proportions of people in groups defined by education and sex to match the

population. If education and sex are correlated with variables or processes of interest in the survey, post-stratification in this way will alter (and presumably improve) the accuracy of the results of statistical analyses. Even social psychologists who rely on others to collect their survey data and to compute their weights should understand how the weights are computed in order to apply them properly during analysis.

Fortunately, social psychologists can benefit from very recent developments that make the computation of weights very easy. This is partly thanks to a blue-ribbon panel of sampling experts who provided advice to the American National Election Studies on how best to implement this sort of computational exercise (see Debell & Krosnick, 2009). Pasek (2012a, 2012b) has written software that is available at no cost for implementation in R that social psychologists can use relatively easily.

Simple Random Sampling.

Simple random sampling is the most basic form of probability sampling. With this method, elements are drawn from the population at random, and all elements have the same chance of being selected. Simple random sampling can be done with or without replacement, where replacement refers to returning selected elements to the population, making them eligible to be selected again. In practice, sampling without replacement (i.e., so that each element has the potential to be selected only once) is most common.

Although conceptually a very straightforward procedure, in practice, simple random sampling is rarely done. Doing it requires that the researcher have a list of all members of the population in advance of drawing the sample, so that elements can be independently and directly selected from the full population listing (the sampling frame). The simple random sample is drawn from the frame by applying a series of random numbers that lead to certain elements being chosen and others not. This can be done for a survey of, say, all of the employees of a particular company, and it can be done for surveys of the general population of some nations other than the United States, which maintain and update a list of all citizens and noncitizen residents of their countries, such as the Netherlands. But the United States does not maintain and distribute a list of all people living in the country, so it is not possible to draw a simple random sample from the population of all Americans. Nonetheless, a random sample can be drawn from this population using other, more complex methods, as we explain later.

Systematic Sampling.

Systematic sampling is a variation of simple random sampling that is slightly more convenient to execute (e.g., Ahlgren, 1983; Kim, Scheufele, Shanahan, & Choi, 2011; Wright, Middleton, & Yon, 2012). Like simple random sampling, systematic sampling requires that all elements be identified and listed. Based on the number of elements in the population and the desired sample size, a sampling interval is determined. For example, if a population contains 20,000 elements, and a sample of 2,000 is desired, the appropriate sampling interval would be 10. That is, every 10th element would be selected to arrive at a sample of the desired size.

To start the sampling process in this example, a random number between 1 and 10 is chosen, and the element on the list that corresponds to this number is included in the sample. This randomly selected number is then used as the starting point for choosing all other elements. Say, for example, the randomly selected starting point was 7 in a systematic sample with a sampling interval of 10. The 7th element on the list would be the first to be included in the sample, followed by the 17th element, the 27th element, and so forth.⁴

We can only be confident that systematic sampling will yield a sample that is representative of the sampling frame from which it was drawn if the elements composing the list have been arranged in a random order. When the elements are arranged in some nonrandom pattern, systematic sampling will not necessarily yield samples that are representative of the populations from which they are drawn. This potential problem is exacerbated when the elements are listed in a cyclical pattern. If the cyclical pattern of elements coincided with the sampling interval, one would draw a distinctly unrepresentative sample.

To illustrate this point, consider a researcher interested in drawing a systematic sample of men and women who had sought marital counseling within the last five years. Suppose he or she obtained a sampling frame consisting of a list of individuals meeting this criterion, arranged by couple: each husband's name listed first, followed by the wife's name. If the researcher's randomly chosen sampling interval was an even number, he or she would end up with a sample composed exclusively of women or exclusively of men, depending on the random start value. This problem is referred to as *periodicity*, and it can be easily avoided by randomizing the order of elements within the sampling frame before applying the selection scheme.

Stratified Sampling.

Stratified sampling is a hybrid of random and systematic sampling, where the sampling frame is divided into subgroups (i.e., strata) and the sampling process is executed either separately on each stratum (e.g., Green & Gerber, 2006; Link, Battaglia, Frankel, Osborn, & Mokdad, 2007; Ross, 1988; Stapp & Fulcher, 1983) or systematically across the entire set of strata. In the example mentioned in the preceding subsection, the sampling frame could be divided into categories (e.g., husbands and wives) and elements could be selected from each category by either a random or systematic method. Stratified sampling provides greater control over the composition of the sample, assuring the researcher of representativeness of the sample in terms of the stratification variable(s). That is, the researcher can implement sampling within genders in order to assure that the ratio of husbands to wives in the sample exactly matches the ratio in the population. When the stratification variable is related to the dependent variable of interest, stratified sampling reduces sampling error below what would result from simple random sampling.

Stratification that involves the use of the same sampling fraction in each stratum is referred to as proportional stratified sampling. Disproportional stratified sampling – using different sampling fractions in different strata – can also be done. This is typically done when a researcher is interested in reducing the standard error in a stratum where the standard deviation is expected to be high. By increasing the sampling fraction in that stratum, he or she can increase the number of elements allocated to the stratum. This is often done to ensure large enough subsamples for subpopulation analyses. For example, a researcher might increase the sampling fraction (often called oversampling) for minority groups in a national survey so that reliable parameter estimates can be generated for such subgroups. It is important to bear in mind here that a representative sample is achieved as long as every member of the population has a known, nonzero probability of being selected into the sample, even if different individuals in the population have different selection probabilities. In other words, random sampling does not require that all members of the population have the same probability of being selected.

Stratification requires that researchers know in advance which variables represent meaningful distinctions between elements in the population. In the example presented earlier, gender was known to be an important dimension, and substantive differences were expected to exist between men and women who had sought marital counseling in the past five years. Of course, if gender were

uncorrelated with the dependent variables, it would not matter if the sample included only men or only women. As Kish (1965) pointed out, the magnitude of the advantage of stratification depends on the relation between the stratification variable and the variable(s) of substantive interest in a study; the stronger this relation, the greater the gain in reducing sampling error from using a stratified sampling strategy. This gain is manifested by greater precision of estimates and more confidence in one's conclusions.

Cluster Sampling.

When a population is dispersed over a broad geographic region, simple random sampling and systematic sampling should yield a sample that is also dispersed broadly. This presents a practical (and costly) challenge in conducting face-to-face interviews, because it is expensive and time-consuming to transport interviewers to widely disparate locations, collecting data from only a small number of respondents in any one place.

To avoid this problem, researchers sometimes implement cluster sampling, which involves drawing a sample with elements in groups (“clusters”) rather than one by one (e.g., Roberto & Scott, 1986; Tziner, 1987). Then all elements within a selected cluster are sampled. From the full geographic region of interest, the researcher might randomly select census tracts, for example, and try to collect data from all of the households in each selected neighborhood.

Cluster sampling has another advantage as well: It permits drawing a random sample from a population when a researcher does not have a list of all population members. For example, if a researcher wishes to conduct a survey of a representative sample of all American residents, it is possible to purchase a set of addresses selected from the U.S. Postal Service's list of all blocks in the country. A researcher could then draw a random sample of blocks and interview everyone whose primary residence is on the selected blocks. Because everyone has an equal and known probability of being selected, this approach will yield a random sample of the nation, even though it is clustered. Face-to-face interviewing of the American adult population is typically done in clusters of households within randomly selected neighborhoods, keeping the cost of maintaining and deploying national interviewing staffs at a manageable level.

Cluster sampling can also be implemented in multiple stages, with two or more sequential steps of random sampling; this is called *multistage* sampling (e.g., Himmelfarb & Norris, 1987; Li, 2008; Shen, Wang, Guo, & Guo, 2009). To assemble a national sample for an in-person survey, for example, one might

begin by randomly selecting 100 or so counties from among the more than 3,000 in the nation. Within each selected county, one could then randomly select a census tract, and from each selected tract one could select a specific census block or its equivalent. Then a certain number of households on each selected block could be randomly selected for inclusion in the sample. To do this, a researcher would need a list of all counties in the United States, all of the census tracts in the selected counties, and all the blocks within the selected tracts, and only then would one need to enumerate all of the households on the selected blocks, from which to finally draw the sample elements. That is, the researcher need not begin with a complete listing of all members of the population.

Cluster sampling can substantially reduce the time and cost of face-to-face data collection, but it also reduces accuracy by increasing sampling error. Members of a cluster are likely to share not only proximity but other attributes as well; they are likely to be more similar to one another along many dimensions than a sample of randomly selected individuals would be. Therefore, interviews with a cluster of respondents will typically yield less precise information about the full population than would the same number of interviews with randomly selected individuals. For statistical tests to be unbiased, this sort of nonindependence needs to be statistically modeled and incorporated in any analysis, thus making the enterprise more cumbersome.

Typical Sampling Methods.

In practice, each mode of survey data collection has its most popular sampling method. Face-to-face surveys of geographically distributed probability samples (e.g., of all Americans) typically involve multistage cluster sampling. In recent years, this method applied to households in the United States begins by purchasing a list of addresses based on the Delivery Sequence File (DSF) assembled by the U.S. Postal Service, and drawing a sample from it. Random Digit Dial (RDD) telephone surveys typically begin with all working area codes and all working central office codes (the next three digits after the area code) for landlines and cell phones and attach four randomly generated digits to yield a random sample of phone numbers. For mail surveys of general population samples, researchers can also begin with a list of addresses, perhaps one that is based on the U.S. Postal Service's DSF, and can draw a random sample from it.

With the spread of Internet access around the world, survey research firms have shifted a great deal of their data collection for academic and industry clients to Internet surveys. And it is possible to conduct such surveys with

probability samples recruited either face-to-face, by telephone, or via mailed invitations. This notion was pioneered by Willem Saris (1998) in the Netherlands, who placed computers and telephone modems in the homes of a random sample of Dutch residents. They completed survey questionnaires regularly via their computers and modems. Saris drew his sample from the Dutch government's list of all residents of the country.

This idea was transported to the United States by a firm called Knowledge Networks (now called GfK), who recruited panels of people to complete surveys regularly via the Internet. Recruitment was originally done by Random Digit Dialing telephone calls, and in more recent years, some recruitment has been done via mailed paper-and-pencil invitations. Computer equipment and Internet access have been provided to all participating individuals who lacked either. This panel has produced remarkably accurate measurements (Chang & Krosnick, 2009; Yeager et al., 2011).

Threats to Sample Representativeness

Ideally, these sampling processes will yield samples that are perfectly representative of the populations from which they were drawn. In practice, however, this virtually never occurs. Sampling error, nonresponse error, and coverage error can distort survey results by compromising representativeness, and social psychologists should understand how this can occur, so as to understand how it can be taken into account during data analysis.

Sampling Error.

Sampling error refers to the discrepancies between values computed from the sample data (e.g., sample means) and the true population values. Such discrepancies are attributable to random differences between the initially chosen sample and the sampling frame from which the sample is drawn. When one uses a probability sample, estimates of the amount of sampling error can be calculated, representing the magnitude of uncertainty regarding obtained parameter estimates resulting from the fact that only a sample from the population was interviewed. Sampling error is typically expressed in terms of the standard error of an estimate, which refers to the variability of sample estimates around the true population value, assuming repeated sampling. That is, the standard error indicates the probability of observing sample estimates of varying distances from the true population value, assuming that an infinite number of samples of a particular size are drawn simultaneously from the same population.

Probability theory provides an equation for calculating the standard error for a single sample from a population of “infinite” size:

$$SE = \sqrt{\text{sample variance/sample size.}} \quad (16.1)$$

With a probability sample, once the standard error has been calculated, it can be used to construct a confidence interval around a sample estimate, which is informative regarding the precision of the parameter estimate. For example, a researcher can be 95% confident that the true population parameter value (e.g., the population's mean value on some variable) falls in the interval that is within 1.96 standard errors of the observed statistic generated from a large sample. A small standard error, then, suggests that the sample statistic provides a relatively precise estimate of the population parameter.

As Equation 16.1 shows, one determinant of sampling error is sample size – as sample size increases, sampling error decreases. This decrease is not linear, however. Moving from a small (e.g., 100) to a moderate sample size (e.g., 500) produces a substantial decrease in sampling error, but further increases in sample size produce smaller and smaller decrements in sampling error. Thus, researchers are faced with a trade-off between the considerable costs associated with increases in sample size and the small relative gains such increases often afford in accuracy.

The formula in Equation 16.1 is correct only if the population size is infinite. When the population is finite, a correction factor may need to be added to the formula for the standard error. Thus, the ratio of sample size to population size is another determinant of sampling error. Data collected from 500 people will include more sampling error if the sample was drawn from a population of 100,000 people than if the sample was drawn from a population of only 1,000 people. When sampling from relatively small populations (i.e., when the sample to population ratio is high), the following alternative sampling error formula should be used:

$$SE = \sqrt{\left(\frac{\text{sample variance}}{\text{sample size}}\right) \left(\frac{\text{population size} - \text{sample size}}{\text{population size}}\right)} \quad (16.2)$$

As a general rule of thumb, this correction only needs to be done when the

sample contains more than 5% of the population (Henry, 1990). However, even major differences in the ratio of the sample size to population size have only a minor impact on sampling error. For example, if a dichotomous variable has a 50/50 distribution in the population and a sample of 1,000 elements is drawn, the standard sampling error formula would lead to a confidence interval of approximately 6 percentage points in width. If the population were only 1,500 in size (i.e., two-thirds of the elements were sampled), the confidence interval width would be reduced to 5 percentage points.

As Equations 16.1 and 16.2 illustrate, sampling error is also dependent on the amount of variance in the variable of interest. If there is no variance in the variable of interest, a sample of one is sufficient to estimate the population value with no sampling error. And as the variance increases, sampling error also increases. With a sample of 1,000, the distribution of a dichotomous variable with a 50/50 distribution in the population can be estimated with a confidence interval 6 percentage points in width. However, the distribution of a dichotomous variable with a 10/90 distribution would have a confidence interval of approximately 3.7 percentage points in width.

The standard formula for calculating sampling error, used by most computer statistical programs, is based on the assumption that the sample was drawn using simple random sampling. When another probability sampling method has been used, the sampling error may actually be slightly higher or slightly lower than the standard formula indicates. This impact of sampling strategy on sampling error is called a *design effect* (deff). Defined more formally, the design effect associated with a probability sample is “the ratio of the actual variance of a sample to the variance of a simple random sample of the same elements” (Kish, 1965, p. 258).

Any probability sampling design that uses clustering will have a design effect in excess of 1.0. That is, the sampling error for cluster sampling will be higher than the sampling error for simple random sampling. Any stratified sampling design, on the other hand, will have a design effect less than 1.0, indicating that the sampling error is lower for stratified samples than for simple random samples. The degree to which the design effect is less than 1.0 depends on the degree to which the stratification variable is related to the outcome variable. Social psychologists should be attentive to design effects, because taking them into account can increase the likelihood of statistical tests detecting genuinely significant effects.

Nonresponse Error.

Even when probability sampling is done for a survey, it is unlikely that 100% of the sampled elements will be successfully contacted and will agree to provide data. Therefore, almost all survey samples include some elements from whom no data were gathered.⁵ A survey's findings may be subject to nonresponse error to the extent that the sampled elements from whom no data were gathered differ systematically and in nonnegligible ways from those from whom data were gathered.

To minimize the potential for nonresponse error, researchers have traditionally implemented various procedures to encourage as many selected respondents as possible to participate (e.g., Dillman, 1978; Fowler, 1988; Lavrakas, 2010). Stated generally, the goal here is to minimize the apparent costs of responding, maximize the apparent rewards for doing so, and establish trust that those rewards will be delivered (Dillman, 1978). One concrete approach to accomplishing these goals is sending “advance” letters to potential respondents informing them that they have been selected to participate in a study and will soon be contacted to do so, explaining that their participation is essential for the study's success because of their expertise on the topic, suggesting reasons why participation will be enjoyable and worthwhile, assuring respondents of confidentiality, and informing them of the study's purpose and its sponsor's credibility. Researchers also make numerous attempts to contact hard-to-reach people and to convince reluctant respondents to participate and sometimes pay people for participation or give them gifts as inducements (e.g., movie passes, pens, golf balls). Such material incentives are effective at increasing participation rates, especially when they are provided at the time the participation invitation is offered, rather than if they are promised to be provided after the interview is completed (e.g., Cantor, O'Hare, & O'Connor, 2008; Singer, Van Howeyk, & Maher, 2000; see Singer & Ye, 2013 for a review on incentives in surveys).

In even the best surveys with the best response rates, there are usually significant biases in the demographic composition of samples. For example, Brehm (1993) showed that in the two leading, recurring academic national surveys of public opinion (the National Election Studies and the General Social Surveys), certain demographic groups were routinely represented in misleading numbers. For example, young adults and old adults are underrepresented, males are underrepresented, people with the highest levels of education are overrepresented, and people with the highest incomes are underrepresented.

Likewise, Smith (1983) reported evidence suggesting that people who do not participate in surveys are likely to have a number of distinguishing demographic characteristics (e.g., living in big cities and working long hours). Holbrook, Krosnick, and Pfent (2008) reported similar evidence.

In most cases, the farther a survey's response rate falls below 100%, the more a researcher can justify concern about the representativeness of the participating sample of respondents. That is, in general, as a survey's response rate drops, the risk of the so-called nonresponse error rises. In other words, the participating sample may be systematically different from the population if nonrespondents are not a random subset of the population. This point of view was expressed especially clearly in a relatively recent revision of guidelines issued by the U.S. Office of Management and Budget regarding procedures for conducted federal surveys in America (Office of Information and Regulatory Affairs, 2006).

However, a high rate of nonresponse does not necessarily mean that a study's measurements of nondemographic variables are fraught with error (c f. Groves, 2006). If the constructs of interest are not correlated with the likelihood of participation, then nonresponse would not distort results. So investing large amounts of money and staff effort to increase response rates might not translate into higher data quality.

A particularly dramatic demonstration of this fact was reported by Visser, Krosnick, Marquette, and Curtin (1996). These researchers compared the accuracy of self-administered mail surveys and telephone surveys forecasting the outcomes of statewide elections in Ohio over a 15-year period. Although the mail surveys had response rates of about 20% and the telephone surveys had response rates of about 60%, the mail surveys predicted election outcomes much more accurately (average error = 1.6%) than the telephone surveys did (average error = 5.2%). In addition, the mail surveys documented the demographic characteristics of voters more accurately than did the telephone surveys. Therefore, simply having a low response rate does not necessarily mean that a survey suffers from a large amount of nonresponse error.

Other studies exploring the impact of response rates have also supported the same conclusion. For example, Brehm (1993) found that statistically correcting for demographic biases in sample composition had very little impact on the substantive implications of correlational analyses. Holbrook *et al.* (2008) meta-analyzed a large set of telephone surveys with widely varying response rates and found that the accuracy of the samples in describing the population declined only very slightly as the response rate fell. Curtin, Presser, and Singer (2000)

reanalyzed a survey dataset to see how much the substantive conclusions of the research differed if the researchers discarded more and more data to simulate determine what the survey's results would have been if the response rate had been lower and lower because the researchers terminated interviews earlier and earlier in the survey's field period. The results changed remarkably little.

In another study, Keeter *et al.* (2000) conducted two simultaneous surveys using the same questionnaire, one employing procedures to increase the response rate as much as possible, and the other taking few such steps. As expected, the former survey yielded a notably higher response rate than did the latter. But the substantive results of the two surveys differed little from one another. Lastly, Merkle and Edelman (2002) analyzed data from exit polls conducted on election day, wherein the response rate for interviewing varied widely from precinct to precinct. The response rate was essentially uncorrelated with the accuracy of the survey's measurement of voting in each precinct. Thus, an accumulating number of publications investigating a wide range of measures show that as long as a random sample is scientifically drawn from the population and thorough, professional efforts are made to collect data from all selected potential respondents, a substantial increase in a survey's response rate is not associated with a notable increase in the accuracy of the survey's results (but see Traugott, Groves, & Lepkowski, 1987).

Nonetheless, it is worthwhile to assess the degree to which nonresponse error is likely to have biased data from any particular sample of interest. One approach to doing so involves making aggressive efforts to recontact a randomly selected sample of people who refused to participate in the survey and collect some data from these individuals. One would especially want to collect data on the key variables of interest in the study, but it can also be useful to collect data on those dimensions along which nonrespondents and respondents seem most likely to differ substantially (Brehm, 1993). A researcher is then in a position to assess the magnitude of differences between people who agreed to participate in the survey and those who refused to do so.

A second strategy rests on the assumption that respondents from whom data were difficult to obtain (either because they were difficult to reach or because they initially declined to participate and were later persuaded to do so) are likely to be more similar to nonrespondents than are people from whom data were relatively easy to obtain. Researchers can compare responses of people who were immediately willing to participate with those of people who had to be recontacted and persuaded to participate. The smaller the discrepancies between

these groups, the less of a threat nonresponse error would seem to be (but see Lin & Schaeffer, 1995 and Mazza & Enders, Chapter 24 in this volume) .

Coverage Error.

One other possible error deserves mention: coverage error. For reasons of economy, researchers sometimes draw probability samples not from the full set of elements in a population of interest but rather from more limited sampling frames. The greater the discrepancy between the population and the sampling frame, the greater potential for coverage error. Such error may invalidate inferences about the population that are made on the basis of data collected from the sample.

By way of illustration, many national surveys these days involve telephone interviewing. And although their goal is to represent the entire country's population, the sampling methods used restrict the sampling frame to individuals with cell phones or living in households with landline telephones. Although the vast majority of American adults do have cell phones or live in households with working telephones, about 5% of the nation does not at any one time. To the extent that people who cannot be reached by phones are different from the rest of the population, generalization of sample results may be inappropriate.

More strikingly, an increasing number of telephone surveys today are so-called robo-polls, meaning that no human interviewers are involved. Instead, an audio recording of a survey's introduction and questions is made, and computers automatically dial randomly generated telephone numbers and play the recording to prospective respondents, who answer questions by either pushing buttons on touch-tone phones or answering orally, and voice-recognition software is used to interpret and record responses. However, because it is illegal in the United States for computers to automatically dial cell phones, robo-polls involve calls only to landline phones, which causes omission from the survey sample of the substantial portion of Americans who do not have a working landline in their homes (about 40% of adults in the United States in 2012; see Blumberg & Luke, 2012). This constitutes a substantial amount of noncoverage, and individuals without landlines are systematically different from those with landlines. Although some robo-polls have yielded reasonable accuracy in anticipating the results of elections, that accuracy appears likely to be an illusory result of adjusting results to match those of previously released high-quality surveys (Clinton & Rogers, 2012). Thus, the noncoverage bias may have been quite consequential.

Nonprobability Sampling

Social psychologists interested in making statements about the general population must employ probability sampling in order to have a scientific justification for generalization. But most social psychological studies have instead been conducted using nonprobability samples, even in domains that conceptually call for probability samples. For example, nonprobability sampling has been used frequently in studies inspired by the surge of interest among social psychologists in the impact of culture on social and psychological processes (e.g., Kitayama & Markus, 1994; Nisbett & Cohen, 1996). In a spate of articles published in top journals, a sample of people from one country was compared with a sample of people from another country, and differences between the samples were attributed to the impact of the countries' cultures (e.g., Benet & Waller, 1995; Hamamura, Meijer, Heine, Kamaya, & Hori, 2009; Han & Shavitt, 1994; Heine & Lehman, 1995; Kitayama, Park, Sevincer, Karasawa, & Uskul, 2009; Rhee, Uleman, Lee, & Roman, 1995). However, in order to convincingly make such comparisons and properly attribute differences to culture, of course, the sample drawn from each culture must be representative of it. And for this to be so, one of the probability sampling procedures described earlier must be used in each country being compared.

Alternatively, one might assume that cultural impact is so universal within a country that any arbitrary sample of people will reflect it. However, hundreds of studies of Americans have documented numerous variations between subgroups within the culture in social psychological processes, and even recent work on the impact of culture has documented variation within nations (e.g., Graham, Haidt, & Nosek, 2009; Nisbett & Cohen, 1996). Therefore, it is difficult to have much confidence in the presumption that any given social psychological process is universal within any given culture, so probability sampling seems essential to permit a reliable conclusion about differences between cultures based on differences between samples of them.

In this light, it is striking that nearly all recent social psychological studies of culture have employed nonprobability sampling procedures. These are procedures where some elements in the population have a zero probability of being selected or have an unknown probability of being selected. For example, Heine and Lehman (1995) compared college students enrolled in psychology courses in a public and private university in Japan with college students enrolled in a psychology course at a public university in Canada. Rhee *et al.* (1995) compared students enrolled in introductory psychology courses at New York

University with psychology majors at Yonsei University in Seoul, Korea. Han and Shavitt (1994) compared undergraduates at the University of Illinois with students enrolled in introductory communication or advertising classes at a major university in Seoul. And Benet and Waller (1995) compared students enrolled at two universities in Spain with Americans listed in the California Twin Registry.

In all of these studies, the researchers generalized the findings from the samples of each culture to the entire cultures they were presumed to represent. For example, after assessing the extent to which their two samples manifested self-enhancing biases, Heine and Lehman (1995) concluded that “people from cultures representative of an interdependent construal of the self,” instantiated by the Japanese students, “do not self-enhance to the same extent as people from cultures characteristic of an independent self,” instantiated by the Canadian students (p. 605). Yet the method of recruiting potential respondents for these studies rendered zero selection probabilities for large segments of the relevant populations. Consequently, it is impossible to know whether the obtained samples were representative of those populations, and it is impossible to estimate sampling error or to construct confidence intervals for parameter estimates. As a result, the statistical calculations used in these articles to compare the different samples were invalid because they presumed simple random sampling from a frame that covered the entire population of interest. Although the researchers using methods like this might argue that the comparisons are valid because the sampling frame was equivalent in the two countries (e.g., college students taking particular courses), no systematic random sampling was actually done from any representative frame, so generalization beyond the research participants is not justified.

More importantly, their results are open to alternative interpretations, as is illustrated by Benet and Waller's (1995) study. One of the authors' conclusions is that in contrast to Americans, “Spaniards endorse a ‘radical’ form of individualism” (Benet & Waller, 1995, p. 715). Justifying this conclusion, ratings of the terms “unconventional,” “peculiar,” and “odd” loaded in a factor analysis on the same factor as ratings of “admirable” and “high-ranking” in the Spanish sample, but not in the American sample. However, Benet and Waller's American college student sample was significantly younger and more homogeneous in terms of age than their sample of Spaniards (the average ages were 24 years and 37 years, respectively; the standard deviations of ages were 4 years and 16 years, respectively). Among Americans, young adults most likely value unconventionality more than older adults do, so what may appear in this

study to be a difference between countries attributable to culture may instead simply be an effect of age that would be apparent within both cultures.

The nonprobability sampling method used most often in the studies described earlier is called *haphazard sampling*, because respondents were selected solely on the basis of convenience (e.g., because they were enrolled in a particular course at a particular university). In some cases, notices seeking volunteers were widely publicized, and people who contacted the researchers were paid for their participation (e.g., Han & Shavitt, 1994). This is problematic because people who volunteer tend to be more interested in (and sometimes more knowledgeable about) the survey topic than the general public (e.g., Bogaert, 1996; Coye, 1985; Dollinger & Leong, 1993), and social psychological processes seem likely to vary with interest and expertise.

Yet another nonprobability sampling method is *purposive sampling*, which involves haphazardly selecting members of a particular subgroup within a population. This technique has been used in a number of social psychological studies to afford comparisons of what are called “known groups” (e.g., Hovland, Harvey, & Sherif, 1957; Webster & Kruglanski, 1994). For example, in order to study people strongly supporting prohibition, Hovland *et al.* (1957) recruited respondents from the Women's Christian Temperance Union, students preparing for the ministry, and students enrolled in religious colleges. And to compare people who were high and low in need for closure, Webster and Kruglanski (1994) studied accounting majors and studio art majors, respectively.

In these studies, the groups of respondents did indeed possess the expected characteristics, but they may as well have had other characteristics that may have been responsible for the studies' results. This is so because the selection procedures used typically yield unusual homogeneity within the “known groups” in at least some regards and perhaps many. For example, accounting majors may have more training in mathematics and related thinking styles than studio art majors do. Had more heterogeneous groups of people high and low in need for closure been studied by Webster and Kruglanski (1994), it is less likely that they would have sharply differed in other regards and less likely that such factors could provide alternative explanations for the results observed.

Snowball sampling is a variant of purposive sampling, where a few members of a rare subpopulation are located, and each is asked to suggest other members of the subpopulation for the researcher to contact. Judd and Johnson (1981) used this method in an investigation comparing people with extreme views on women's issues to people with moderate views. To assemble a sample of people

with extreme views, these investigators initially contacted undergraduate women who were members of feminist organizations and then asked them to provide names of other women who were also likely to hold similar views on women's issues. Like cluster sampling, this sampling method also violates the assumption of independence of observations, complicating analysis. Recent developments with a related technique called “respondent-driven sampling” have sought to systematize the application of snowball sampling (Heckathorn, 1997, 2002; Salganik & Heckathorn, 2004).

Probably the best-known form of nonprobability sampling is *quota sampling*, which involves selecting members of various subgroups of the population to build a sample that accurately reflects certain known characteristics of the population. Predetermined numbers of people in each of several categories are recruited to accomplish this. For example, one can set out to recruit a sample half of which is comprised of men and another half of women, and one-third of people with less than high school education, one-third of people with only a high school degree, and one-third of people with at least some college education.

If quotas are imposed on a probability sampling procedure (e.g., telephone interviews done by random digit dialing) and if the quotas are based on accurate information about a population's composition (e.g., the U.S. Census), then the resulting sample may be more accurate than simple random sampling would be, although the gain most likely would be very small.

However, quotas are not usually imposed on probability sampling procedures but instead are imposed on haphazard samples. Therefore, this approach can give an arbitrary sample the patina of representativeness, when in fact only the distributions of the quota criteria match the population. A particularly dramatic illustration of this problem is the failure of pre-election polls to predict that Truman would win his bid for the U.S. presidency in 1948. Although interviewers conformed to certain demographic quotas in selecting respondents, the resulting sample was quite unrepresentative in some regards not explicitly addressed by the quotas (Mosteller, Hyman, McCarthy, Marks, & Truman, 1949). A study by Katz (1942) illustrated how interviewers tend to oversample residents of one-family houses, American-born people, and well-educated people when these dimensions are not explicit among the quota criteria.

Although surveys done with probability samples and collecting data via the Internet yield remarkably accurate results (Chang & Krosnick, 2009; Yeager et al., 2011), the vast majority of Internet survey data being collected around the world is based on nonprobability “volunteer” (so called opt-in) samples of

respondents. This is an especially surprising development from a scientific standpoint, because survey professionals learned their lesson decades ago about the dangers of nonprobability sampling. Not only does this approach lack theoretical foundation, but it yields findings that are consistently less accurate than results produced with probability samples (e.g., Chang & Krosnick, 2009; Yeager et al., 2011). Despite claims to the contrary made by most of the companies that collect and sell such nonprobability sample Internet data, their methods appear not to be as accurate as those produced by probability samples.

Given all this, we urge researchers to recognize the inherent limitations of nonprobability sampling methods and to draw conclusions about populations or about differences between populations tentatively, if at all, when nonprobability sampling methods are used. Furthermore, we encourage researchers to attempt to assess the representativeness of samples they study by comparing their attributes with known population attributes in order to bolster confidence in generalization when appropriate and to temper such confidence when necessary.

Are we suggesting that all studies of college sophomores enrolled in introductory psychology courses are of minimal scientific value? Absolutely not. The value of the vast majority of social psychological laboratory experiments does not hinge on generalizing their results to a population. Instead, these studies test whether a particular process occurs at all, to explore its mechanisms, and to identify its moderators. Any demonstrations along these lines enhance our understanding of the human mind, even if the phenomena documented occur only among select groups of American college sophomores.

After an initial demonstration of an effect, process, or tendency, subsequent research can assess its generality. Therefore, work such as Heine and Lehman's (1995) is valuable because it shows us that some findings are not limitlessly generalizable and sets the stage for research illuminating the relevant limiting conditions. We must be careful, however, about presuming that we know what these limiting conditions are without proper, direct, and compelling tests of our conjectures.

Questionnaire Design and Measurement Error

Once a sample is selected, the next step for a survey researcher is questionnaire design. When designing a questionnaire, a series of decisions must be made about each question. First, will it be open-ended or closed-ended? And for some closed-ended question tasks, should one use rating scales or ranking tasks? If

one uses rating scales, how many points should be on the scales and how should they be labeled with words? Should respondents be explicitly offered “no-opinion” response options or should these be omitted? In what order should response alternatives be offered? How should question stems be worded? And finally, once all the questions are written, decisions must be made about the order in which they will be asked.

Every researcher's goal is to maximize the reliability and validity of the data he or she collects. Therefore, each of the aforementioned design decisions should presumably be made so as to maximize these two indicators of data quality. Fortunately, thousands of empirical studies provide clear and surprisingly unanimous advice on the issues listed in the preceding paragraph. Although a detailed review of this literature is beyond the scope of this chapter (for reviews, see Bradburn, et al., 1981; Converse & Presser, 1986; Krosnick & Fabrigar, forthcoming; Saris & Gallhofer, 2007; Schuman & Presser, 1981; Sudman, Bradburn, & Schwarz, 1996), we provide a brief tour of the implications of these studies. John and Benet-Martinez (Chapter 18 in this volume) discuss reliability in more detail; Brewer and Crano (Chapter 2 in this volume) discuss validity.

Open vs. Closed Questions

An open-ended question permits the respondent to answer in his or her own words. For example, in political surveys one commonly asked nominal open-ended question is “What is the most important problem facing the country today?” In contrast, a closed-ended question requires that the respondent select an answer from a set of choices offered explicitly by the researcher. A closed-ended version of the above question might ask: “What is the most important facing the country today: inflation, unemployment, crime, the federal budget deficit, or some other problem?”

The biggest challenge in using open-ended questions is the task of coding responses. In a survey of 1,000 respondents, nearly 1,000 different verbatim answers will be given to the “most important problem” question if considered word for word. But in order to analyze these answers statistically, they must be clumped into a relatively small number of categories. This requires that a set of mutually exclusive and exhaustive codes be developed for each open-ended question. Multiple people should read and code the answers into the categories, the level of agreement between the coders must be ascertained, and the procedure must be refined and repeated if agreement is too low. The time and

financial costs of such a procedure, coupled with the added challenge of requiring interviewers to carefully transcribe answers, have led many researchers to favor closed-ended questions, which in essence ask respondents to directly code themselves into categories that the researcher specifies.

Unfortunately, when used in certain applications, closed-ended questions have distinct disadvantages. Most importantly, respondents tend to confine their answers to the choices offered, even if the researcher does not wish them to do so (Jenkins, 1935; Lindzey & Guest, 1951; Presser, 1990). Explicitly offering the option to specify a different response does little to combat this problem. If the list of choices offered by a question is incomplete, even the rank ordering of the choices that are explicitly offered can be different from what would be obtained from an open-ended question. Therefore, a closed-ended question can only be used effectively if its answer choices are comprehensive, and this can often be assured only if an open-ended version of the question is administered in a pretest using a reasonably large sample. Perhaps, then, researchers should simply include the open-ended question in the final questionnaire because they will otherwise have to deal with the challenges of coding during pretesting. Also supportive of this conclusion is evidence that open-ended questions have higher reliabilities and validities than closed-ended questions (e.g., Haddock & Zanna, 1998; Hurd, 1932; Remmers, Marschat, Brown, & Chapman, 1923; Schuman, 2008; see also Smyth, Dillman, Christian, & McBride, 2009 on open-ended questions in Web surveys).

One might hesitate in implementing this advice because nominal open-ended questions may themselves be susceptible to unique problems. For example, some researchers feared that open-ended questions would not work well for respondents who are not especially articulate, because they might have special difficulty describing their thoughts, opinions, and feelings. However, this seems not to be a problem (England, 1948; Haddock & Zanna, 1998; Geer, 1988). Second, some researchers feared that respondents would be especially likely to answer open-ended questions by mentioning the most salient possible responses, not those that are truly most appropriate. But this, too, appears not to be the case (Schuman, Ludwig, & Krosnick, 1986). Thus, open-ended questions seem to be worth the trouble they take to measure nominal constructs.

Another type of open-ended question seeks a number, such as the number of times that a respondent went out to the movies during the last month. Such a question can also be asked in a closed-ended format, offering ranges. For example, respondents can be asked whether they never went out to the movies,

went out once or twice, or went out three or more times. Offering ranges like this might seem to simply the respondent's task by allowing him or her to answer approximately rather than exactly. But in fact, answering such a question accurately first requires the respondent to answer the open-ended version of the question in his or her own mind and then match that answer to one of the offered ranges. Thus, it would be simpler for respondents to skip the matching step and simply report the answer to the open-ended question. And the particular ranges offered by researchers are usually relatively arbitrarily chosen, yet they can manipulate respondents' answers (Courneya, Jones, Rhodes, & Blanchard, 2003; Hurd, 1999; Richardson, 2004; Schwarz, Hippler, Deutsch, & Strack, 1985; Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991). For example, if a question asks respondents how many hours he/she typically watches television during a week and offers a series of answer choices (e.g., "less than 2 hours, 3–5 hours, 6–8 hours, 9–11 hours, 12 or more hours), respondents infer that the researcher expects to obtain a normal distribution of responses, with the mode in the middle of the range, and infer that the midpoint is the most common behavior in the population. Therefore, respondents gravitate toward the middle of the offer ranges, no matter what ranges are offered. Thus, numeric questions are best asked seeking an exact number from respondents.

Rating versus Ranking

Practical considerations enter into the choice between ranking and rating questions as well. Imagine that one wishes to determine whether people prefer to eat carrots or peas. Respondents could be asked this question directly (a ranking question), or they could be asked to rate their attitudes toward carrots and peas separately, and the researcher could infer which is preferred. With this research goal, asking the single ranking question seems preferable and more direct than asking the two rating questions. But rank-ordering a large set of objects takes longer and is less enjoyed by respondents than a rating task (Elig & Frieze, 1979; Taylor & Kinnear, 1971). Furthermore, ranking might force respondents to make choices between objects toward which they feel identically, and ratings can reveal not only which object a respondent prefers but also how different his or her evaluations of the objects are.

Surprisingly, however, rankings are more effective than ratings, partly because ratings suffer from a significant problem: *nondifferentiation*. When rating a large set of objects on a single scale, a significant number of respondents rate multiple objects identically as a result of *survey satisficing* (Krosnick,

1991b). That is, although these respondents could devote thought to the response task, retrieve relevant information from memory, and report differentiated attitudes toward the objects, they often choose to shortcut this process instead. To do so, they choose what appears to be a reasonable point to rate most objects on the scale and select that point over and over (i.e., nondifferentiation), rather than thinking carefully about each object and rating different objects differently (Krosnick, 1991b; Krosnick & Alwin, 1988). As a result, the reliability and validity of ranking data are superior to those of rating data (e.g., Harzing et al., 2009; Miethe, 1985; Munson & McIntyre, 1979; Nathan & Alexander, 1985; Rankin & Grube, 1980; Reynolds & Jolly, 1980). So although rankings do not yield interval-level measures of the perceived distances between objects in respondents' minds and are more statistically cumbersome to analyze (Alwin & Jackson, 1982), these measures are apparently more useful when a researcher's goal is to ascertain rank orders of objects.

Rating Scale Formats

When designing a rating scale, one must begin by specifying the number of points on the scale. Many studies have compared the reliability and validity of scales of varying lengths (for a review, see Krosnick & Fabrigar, forthcoming). For bipolar scales (e.g., running from positive to negative with neutral in the middle), reliability and validity are highest for about seven points (e.g., Matell & Jacoby, 1971; see also Alwin & Krosnick, 1991; Lozano, Garcia-Cueto, & Muñiz, 2008; Preston & Colman, 2000). In contrast, the reliability and validity of unipolar scales (e.g., running from no importance to very high importance) seem to be optimized for a bit shorter scales, approximately five points long (e.g., Wikman & Warneryd, 1990). Techniques such as magnitude scaling (e.g., Lodge, 1981), which offer scales with an infinite number of points, yield data of lower quality than do more conventional rating scales and should therefore be avoided (e.g., Cooper & Clare, 1981; Miethe, 1985; Patrick, Bush, & Chen, 1973; see also Cook, Heath, & Thompson, 2001 and Couper, Tourangeau, Conrad, & Singer, 2006 on Web-based visual analogue rating scales).

A good number of studies suggest that data quality is better when all scale points are labeled with words than when only some are (e.g., Krosnick & Berent, 1993; Weng, 2004; Weijters, Cabooter, & Schillewaert, 2010). Furthermore, respondents are more satisfied when more rating scale points are verbally labeled (e.g., Dickinson & Zellinger, 1980). Researchers should strive to select labels that have meanings that divide up the continuum into approximately equal

units (e.g., Klockars & Yamagishi, 1988). For example, “very good, good, and poor” is a combination that should be avoided, because the terms do not divide the continuum equally: the meaning of “good” is much closer to the meaning of “very good” than it is to the meaning of “poor” (Myers & Warner, 1968).

Researchers in many fields these days ask people questions offering response choices such as “agree-disagree,” “true-false,” or “yes-no” (e.g., Bearden, Netemeyer, & Mobley, 1993). Yet a great deal of research suggests that these response choices sets are problematic because of acquiescence response bias (e.g., Couch & Keniston, 1960; Jackson, 1979; Schuman & Presser, 1981; Schuman & Scott, 1989). That is, some people are inclined to say “agree,” “true,” or “yes,” regardless of the content of the question. Furthermore, these responses are more common among people with limited cognitive skills, for more difficult items, and for items later in a questionnaire, when respondents are presumably more fatigued (Krosnick, 1991b). A number of studies demonstrate how acquiescence can distort the results of substantive investigations (e.g., Jackman, 1973; Saris, Revilla, Krosnick, & Shaeffer, 2010; Winkler, Kanouse, & Ware, 1982), and in a particularly powerful historical example, acquiescence undermined the scientific value of *The Authoritarian Personality*'s extensive investigation of fascism and anti-Semitism (Adorno, Frankel-Brunswick, Levinson, & Sanford, 1950). This damage occurs equally when dichotomous items offer just two choices (e.g., “agree” and “disagree”) as when a rating scale is used (e.g., ranging from “strongly agree” to “strongly disagree”).

It might seem that acquiescence can be controlled by measuring a construct with a large set of items, half of them making assertions opposite to the other half (called “item reversals”). This approach is designed to place acquiescing responders in the middle of the final dimension but will do so only if the assertions made in the reversals are equally extreme as the statements in the original items. This involves extensive pretesting and is therefore cumbersome to implement. Furthermore, it is difficult to write large sets of item reversals without using the word “not” or other such negations, and evaluating assertions that include negations is cognitively burdensome and error-laden for respondents, thus adding measurement error and increasing respondent fatigue (e.g., Eifermann, 1961; Wason, 1961). And even after all this, acquiescing respondents presumably end up at the midpoint of the resulting measurement dimension, which is probably not where most belong on substantive grounds anyway. That is, if these individuals were induced not to acquiesce but to answer the items thoughtfully, their final index scores would presumably be more valid than placing them at the midpoint.

Most important, answering an agree-disagree, true-false, or yes-no question always involves first answering a comparable rating question in one's mind. For example, if a man is asked to agree or disagree with the assertion "I am not a friendly person," he must first decide how friendly he is (perhaps concluding "very friendly") and then translate that conclusion into the appropriate selection in order to answer the question he was asked ("disagree" to the original item). It would be simpler and more direct to ask the person how friendly he is. In fact, every agree-disagree, true-false, or yes-no question implicitly requires the respondent to make a mental rating of an object along a continuous dimension, so asking about that dimension is simpler, more direct, and less burdensome. It is not surprising, then, that the reliability and validity of other rating scale and forced choice questions are higher than those of agree-disagree, true-false, and yes-no questions (e.g., Ebel, 1982; Mirowsky & Ross, 1991; Ruch & DeGraff, 1926; Wesman, 1946). Consequently, it seems best to avoid long batteries of questions in these latter formats and instead ask just two or three questions using other rating scales and forced choice formats (e.g., Robins, Hendin, & Trzesniewski, 2001) .

The Order of Response Alternatives

The answers people give to closed-ended questions are sometimes influenced by the order in which the alternatives are offered. When categorical response choices are presented visually, as in self-administered questionnaires, people are inclined toward primacy effects, whereby they tend to select answer choices offered early in a list (e.g., Galesic, Tourangeau, Couper, & Conrad, 2008; Krosnick & Alwin, 1987; Miller & Krosnick, 1998; Sudman et al., 1996). But when categorical answer choices are read aloud to people, recency effects tend to appear, whereby people are inclined to select the options offered last (e.g., Holbrook, Krosnick, Moore, & Tourangeau, 2007; McClendon, 1991). And when rating scales are presented visually and orally, primacy effects routinely appear. These effects are most pronounced among respondents low in cognitive skills and when questions are more cognitively demanding (Holbrook et al., 2007; Krosnick & Alwin, 1987; Payne, 1949/1950; Schuman & Presser, 1996). All this is consistent with the theory of satisficing (Krosnick, 1991b), which posits that response order effects are generated by the confluence of a confirmatory bias in evaluation, cognitive fatigue, and a bias in memory favoring response choices read aloud most recently. Therefore, it seems best to minimize the difficulty of questions and to rotate the order of response choices across respondents.

No-Opinion Filters and Attitude Strength

Concerned about the possibility that respondents may feel pressure to offer opinions on issues when they truly have no attitudes (e.g., P. E. Converse, 1964), questionnaire designers have often explicitly offered respondents the option to say they have no opinion. And indeed, many more people say they “don't know” what their opinion is when this is done than when it is not (e.g., Schuman & Presser, 1981; Schuman & Scott, 1989). People tend to offer this response under conditions that seem sensible (e.g., when they lack knowledge on the issue; Donovan & Leivers, 1993; Luskin & Bullock, 2011), and people prefer to be given this option in questionnaires (Ehrlich, 1964). However, most “don't know” responses stem from conflicting feelings or beliefs (rather than lack of feelings or beliefs all together) and uncertainty about exactly what a question's response alternatives mean or what the question is asking (e.g., Coombs & Coombs, 1976; see also Berinsky, 1999). It is not surprising, then, that the quality of data collected is no higher when a “no opinion” option is offered than when it is not (e.g., Krosnick, Holbrook, Berent, Carson, Hanemann, Kopp, Mitchell, Presser, Ruud, Smith, Moody, Green, & Conaway, 2002; McClendon & Alwin, 1993). That is, people who would have selected this option if offered nonetheless give meaningful opinions when it is not offered.

A better way to accomplish the goal of differentiating “real” opinions from “nonattitudes” is to measure the strength of an attitude using one or more follow-up questions. Krosnick and Petty (1995) proposed that strong attitudes can be defined as those that are resistant to change, are stable over time, and have powerful impact on cognition and action. Many empirical investigations have confirmed that attitudes vary in strength, and the respondent's presumed task when confronting a “don't know” response option is to decide whether his or her attitude is sufficiently weak as to be best described by selecting that option. But because the appropriate cut point along the strength dimension seems exceedingly hard to specify, it would seem preferable to ask people to describe where their attitude falls along the strength continuum.

However, there are many different aspects of attitudes related to their strength that are all somewhat independent of each other (e.g., Krosnick, Boninger, Chuang, Berent, & Carnot, 1993; Wojcieszak, 2012). For example, people can be asked how important the issue is to them personally, or how much they have thought about it, or how certain they are of their opinion, or how knowledgeable they are about it (for details on measuring these and many other dimensions, see Wegener, Downing, Krosnick, & Petty, 1995). Each of these dimensions can

help differentiate attitudes that are crystallized and consequential from those that are not.

Question Wording

The logic of questionnaire-based research requires that all respondents be confronted with the same stimulus (i.e., question), so any differences between people in their responses stem from real differences between the people. But if the meaning of a question is ambiguous, different respondents may interpret it differently and respond to it differently. Therefore, experienced survey researchers advise that questions always avoid ambiguity. They also recommend that wordings be easy for respondents to understand (thereby minimizing fatigue), and this can presumably be done by using short, simple words that are familiar to people. When complex or jargony words must be used, it is best to define them explicitly.

Another standard piece of advice from seasoned surveyors is to avoid so-called double-barreled questions, which actually ask two questions at once. Consider the question, “Do you think that parents and teachers should teach middle school students about birth control options?” If a respondent feels that parents should do such teaching and that teachers should not, there is no comfortable way to say so, because the expected answers are simply “yes” or “no.” Questions of this sort should be decomposed into ones that address the two issues separately.

Sometimes, the particular words used in a question stem can have a big impact on responses. For example, Smith (1987) found that respondents in a national survey were much less positive toward “people on welfare” than toward “the poor.” But Schuman and Presser (1981) found that people reacted equivalently to the concepts of “abortion” and “ending pregnancy,” despite the investigators’ intuition that these concepts would elicit different responses. These investigators also found that more people say that a controversial behavior should be “not allowed” than say it should be “forbidden,” despite the apparent conceptual equivalence of the two phrases. Thus, subtle aspects of question wording can sometimes make a big difference, so researchers should be careful to say exactly what they want to say when wording questions. Unfortunately, however, this literature does not yet offer general guidelines or principles about wording selection.

Question Order

An important goal when ordering questions is to help establish a respondent's comfort and motivation to provide high-quality data. If a questionnaire begins with questions about matters that are highly sensitive or controversial, or that require substantial cognitive effort to answer carefully, or that seem poorly written, respondents may become uncomfortable, uninterested, or unmotivated and may therefore terminate their participation. Seasoned questionnaire designers advise beginning with items that are easy to understand and answer on engaging, noncontroversial topics.

Once into a questionnaire a bit, grouping questions by topic may be useful. That is, once a respondent starts thinking about a particular topic, it is presumably easier for him or her to continue to do so, rather than having to switch back and forth between topics, question by question. However, initial questions in a sequence can influence responses to later, related questions, for a variety of reasons (McFarland, 1981; Moore, 2002; Sudman et al., 1996; Tourangeau & Rasinski, 1988; Tourangeau, Rips, & Rasinski, 2000; Van de Walle & Van Ryzin, 2011; Wilson, 2010). For example, when national survey respondents were asked whether it should be possible for a married woman to obtain a legal abortion if she does not want any more children, fewer respondents expressed support after first being asked whether abortion should be legal if there is a strong chance of a serious defect in the baby (Schuman & Presser, 1981). This might be owing to a perceptual contrast effect, given that the latter seems like a more compelling justification than the former is. Also, being asked whether communist news reporters should be allowed to work in the United States in the 1940s, American survey respondents were far more likely answer affirmatively if they had previously been asked if American news reporters should be allowed to work in the Soviet Union (Schuman & Presser, 1981). This may be the result of activation of the “norm of reciprocity” at the time the second question is asked: a norm that states all parties in a dispute should be treated equally. Therefore, within blocks of related questions on a single topic, it might be useful to rotate question order across respondents so that any question order effects can be empirically gauged and statistically controlled for if necessary.

Questions to Avoid

It is often of interest to researchers to study trends over time in attitudes or beliefs. To do so usually requires measuring a construct at repeated time points in the same group of respondents. An appealing shortcut is to ask people to

attempt to recall the attitudes or beliefs they held at specific points in the past. However, a great deal of evidence suggests that people are quite poor at such recall, usually presuming that they have always believed what they believe at the moment (e.g., Bem & McConnell, 1970; Ross, 1989). Therefore, such questions vastly underestimate change and should be avoided unless the researcher wishes to measure people's perceptions of change per se.

Because researchers are often interested in identifying the causes of people's thoughts and actions, it is tempting to ask people directly why they thought a certain thing or behaved in a certain way. This involves asking people to introspect and describe their own cognitive processes, which was one of modern psychology's first core research methods (Hothersall, 1984). However, it became clear to researchers in the 20th century that it did not work well, and Nisbett and Wilson (1977) articulated an argument about why this is so. Evidence produced since their landmark work has largely reinforced the conclusion that many cognitive processes occur very quickly and automatically “behind a black curtain” in people's minds, so they are unaware of them and cannot describe them. Consequently, questions asking for such descriptions seem best avoided as well, unless researchers wish to measure people's perceptions of the causes of their thinking and action per se .

Pretesting

Even the most carefully designed questionnaires sometimes include items that respondents find ambiguous or difficult to comprehend. Questionnaires may also include items that respondents understand perfectly well but interpret differently than the researcher intended. Because of this, questionnaire pretesting is conducted to detect and repair such problems. Pretesting can also provide information about probable response rates of a survey, the cost and time frame of the data collection, the effectiveness of the field organization, and the skill level of the data collection staff. A number of pretesting methods have been developed, each of which has advantages and disadvantages, as we review next.

Pretesting Methods for Interviewer-Administered Questionnaires

Pretesting questionnaires is routine in survey research, but may be done more rarely by social psychologists. Often, questionnaires designed based on intuition or tradition are deployed without determining whether they are effective measuring tools. If social psychologists wish to learn from survey researchers about how to evaluate their questionnaires before deploying them, a variety of techniques are available, as we outline next.

Conventional Pretesting.

In conventional face-to-face and telephone survey pretesting, interviewers conduct a small number of interviews (usually between 15 and 25) and then discuss their experiences with the researcher in a debriefing session (e.g., Bischooping, 1989; Nelson, 1985). They describe any problems they encountered (e.g., identifying questions that required further explanation, wording that was difficult to read or that respondents seemed to find confusing) and their impressions of the respondents' experiences in answering the questions. Researchers might also look for excessive item nonresponse in the pretest interviews, which might suggest a question is problematic. On the basis of this information, researchers can modify the survey instrument to increase the likelihood that the meaning of each item is clear to respondents and that the interviews proceed smoothly.

Conventional pretesting can provide valuable information about the survey

instrument, especially when the interviewers are experienced survey data collectors. But this approach has limitations. For example, what constitutes a “problem” in the survey interview is often defined rather loosely, so there is potential for considerable variance across interviewers in terms of what is reported during debriefing sessions. Also, debriefing interviews are sometimes relatively unstructured, which might further contribute to variance in interviewers’ reports. Of course, researchers can standardize their debriefing interviews, thereby reducing the idiosyncrasies in the reports from pretest interviewers. Nonetheless, interviewers’ impressions of respondent reactions are unavoidably subjective and are likely to be imprecise indicators of the degree to which respondents actually had difficulty with the survey instrument.

Behavior Coding.

A second method, called behavior coding, offers a more objective, standardized approach to pretesting. Behavior coding involves monitoring pretest interviews (either as they take place or via video recordings) and noting events that occur during interactions between the interviewer and the respondent (e.g., Cannell, Miller, & Oksenberg, 1981; Hess, Singer, & Bushery, 1999). The coding reflects each deviation from the script (caused by the interviewer misreading the questionnaire, for example, or by the respondent asking for additional information or providing an initial response that was not sufficiently clear or complete). Questions that elicit frequent deviations from the script are presumed to require modification.

Although behavior coding provides a more systematic, objective approach than conventional pretest methods, it is also subject to limitations. Most important, behavior coding is likely to miss problems centering around misconstrued survey items, which may not elicit any deviations from the script.

Cognitive Interviewing.

To overcome this important weakness, researchers employ a third pretest method, borrowed from cognitive psychology. It involves administering a questionnaire to a small number of people who are asked to “think aloud,” verbalizing whatever considerations come to mind as they formulate their responses (e.g., Beatty & Willis, 2007; Forsyth & Lessler, 1991; Willis, 2005). This “think aloud” procedure is designed to assess the cognitive processes by which respondents answer questions, which presumably provides insight into the way each item is comprehended and the strategies used to devise answers.

Interviewers might also ask respondents about particular elements of a survey question, such as interpretations of a specific word or phrase or overall impressions of what a question was designed to assess.

Comparing these Pretesting Methods.

These three methods of pretesting focus on different aspects of the survey data collection process, and one might expect that they would detect different types of interview problems. And indeed, empirical evidence suggests that the methods do differ in terms of the kinds of problems they detect, as well as in the reliability with which they detect these problems (i.e., the degree to which repeated pretesting of a particular questionnaire consistently detects the same problems).

Presser and Blair (1994) demonstrated that behavior coding is quite consistent in detecting apparent respondent difficulties and interviewer problems. Conventional pretesting also detects both sorts of potential problems, but less reliably. In fact, the correlation between the apparent problems diagnosed in independent conventional pretesting trials of the same questionnaire can be remarkably low. Cognitive interviews also tend to exhibit low reliability across trials, and they tend to detect respondent difficulties almost exclusively.

However, the relative reliability of the various pretesting methods is not necessarily informative about the validity of the insights gained from them. And one might even imagine that low reliability actually reflects the capacity of a particular method to continue to reveal additional, equally valid problems across pretesting iterations. But unfortunately, we know of no empirical studies evaluating or comparing the validity of the various pretesting methods. Much research along these lines is clearly needed (for a review, see Presser, Rothgeb, Couper, Lessler, Martin, Martin, & Singer, 2004) .

Self-Administered Questionnaire Pretesting

Pretesting is especially important when data are to be collected via self-administered questionnaires, because interviewers will not be available to clarify question meaning or probe incomplete answers. Furthermore, with self-administered questionnaires, the researcher must be as concerned about the layout of the questionnaire as with the content; that is, the format must be “user-friendly” for the respondent. Achieving this goal is a particular challenge when doing surveys via the Internet, because different browsers display text and

graphics differently to different respondents (Couper, 2008). A questionnaire that is easy to use can presumably reduce measurement error and may also reduce the potential for nonresponse error by providing a relatively pleasant task for the respondent.

Unfortunately, however, pretesting is also most difficult when self-administered questionnaires are used, because problems with item comprehension or response selection are less evident in self-administered questionnaires than face-to-face or telephone interviews. Some researchers rely on observations of how pretest respondents fill out a questionnaire to infer problems in the instrument – an approach analogous to behavior coding in face-to-face or telephone interviewing. But this is a less than optimal means of detecting weaknesses in the questionnaire.

A more effective way to pretest self-administered questionnaires is to conduct face-to-face interviews with a group of survey respondents drawn from the target population. Researchers can use the previously described “think aloud” procedure, asking respondents to verbalize their thoughts as they complete the questionnaire. Alternatively, respondents can be asked to complete the questionnaire just as they would during actual data collection, after which they can be interviewed about the experience. They can be asked about the clarity of the instructions, the question wording, and the response options. They can also be asked about their interpretations of the questions or their understanding of the response alternatives and about the ease or difficulty of responding to the various items.

Data Collection

The survey research process culminates in the “field period,” during which the data are collected, and the careful execution of this final step is critical to success. Next, we discuss considerations relevant to data collection mode (face-to-face, telephone, and self-administered) and interviewer selection, training, and supervision (for comprehensive discussions, see, e.g., Bradburn & Sudman, 1979; Dillman, 1978, 2007; Fowler & Mangione, 1990; Frey, 1989; Lavrakas, 1993).

Mode

Face-to-Face Interviews.

National face-to-face data collection often requires a large staff of well-trained interviewers who visit respondents in their homes. But this mode of data collection is not limited to in-home interviews; face-to-face interviews can be conducted in a laboratory or other locations as well. Whatever the setting, face-to-face interviews involve the oral presentation of survey questions, sometimes with visual aids. For many years, interviewers recorded responses on paper copies of the questionnaire, but now face-to-face interviewers are equipped with laptop or tablet computers, and the entire data collection process is being regulated by computer software.

In computer-assisted personal interviewing (CAPI; see United Nations Economic and Social Commission for Asia and the Pacific, 1999a), interviewers work from a computer screen, on which the questions to be asked appear one by one in the appropriate order. Responses are typed into the computer, and subsequent questions appear instantly on the screen. This system can reduce some types of interviewer error, and it permits researchers to vary the specific questions each respondent is asked based on responses to previous questions. It also makes the incorporation of experimental manipulations into a survey easy, because the manipulations can be incorporated directly into the CAPI program. In addition, this system eliminates the need to enter responses into a computer after the interview has been completed.

Telephone Interviews.

Instead of interviewing respondents in person, researchers sometimes rely on telephone interviewing as their primary mode of data collection, and such interviewing is almost always driven by software presenting questions on computer screens to interviewers. Responses are typed immediately into the computer. So-called computer-assisted telephone interviewing (CATI; United Nations Economic and Social Commission for Asia and the Pacific, 1999b) is the industry standard, and several software packages are available to simplify computer programming.

Self-Administered Questionnaires.

Self-administration is employed when paper questionnaires are mailed or dropped off to individuals at their homes, along with instructions on how to return the completed surveys. Alternatively, people can be intercepted on the street or in other public places and asked to complete a self-administered questionnaire, or such questionnaires can be distributed to large groups of

individuals gathered specifically for the purpose of participating in the survey or for entirely unrelated purposes (e.g., during a class period or at an employee staff meeting). Whatever the method of distribution, this mode of data collection typically requires respondents to complete a written questionnaire and return it to the researcher.

Although paper questionnaire self-administration continues today in prominent contexts (e.g., the exit polls conducted on election days by major news media organizations), computer self-administration is now much more common. Not only are computers used for Internet surveys, but they can be used in face-to-face interviewing and telephone as well. In-home interviewers usually bring laptop or tablet computers with them and can pass those computers to their respondents, who can listen to questions being read aloud to them on headphones and type their answers directly into the computer. And interactive voice response (IVR) technology can be used during telephone interviews, whereby respondents hear prerecorded audio renderings of the questions and type their answers on their telephone keypads. Thus, computer assisted self-administered interviewing (CASAI) and Audio CASAI afford all of the advantages of computerized face-to-face and telephone interviewing, along with many of the advantages of self-administration (for a review of modes, see Groves et al., 2009).

Smartphones and other mobile devices are now also being used to collect survey data. Although small screens pose challenges for the presentation of questions and the recording of answers, mobile devices offer the advantage of allowing researchers to know some respondents' physical locations at the time the questionnaire is completed and allow respondents to take and send real-time photographs to researchers.

Choosing a Mode

Face-to-face interviews, telephone interviews, and self-administered questionnaires each afford certain advantages, and choosing among them requires trade-offs. This choice should be made with several factors in mind, including cost, characteristics of the population, sampling strategy, desired response rate, question format, question content, questionnaire length, length of the data collection period, and availability of facilities.

Cost.

The first factor to be considered when selecting a mode of data collection is cost. Face-to-face interviews of representative samples of the general population are much more expensive than telephone interviews, which are about as expensive as Internet and paper-and-pencil surveys of representative samples of general populations these days.

The Population.

Several characteristics of the population are relevant to selecting a mode of data collection. For example, completion of a self-administered questionnaire requires a basic proficiency in reading and, depending on the response format, writing or computer operation. Thus, this mode of data collection is inappropriate if a non-negligible portion of the population being studied does not meet this minimum literacy proficiency. Motivation is another relevant factor – when researchers suspect that respondents may be unmotivated to participate in a survey, or to read questions carefully, interviewers are typically more effective at eliciting participation than are paper or email invitations. Skilled interviewers can often increase response rates by convincing individuals of the value of the survey and persuading them to participate and provide high-quality data (Cannell, Oksenberg, & Converse, 1977; Groves, Cialdini, & Couper, 1992; Marquis, Cannell, & Laurent, 1972).

Sampling Strategy.

The sampling strategy to be used may sometimes suggest a particular mode of data collection. For example, some preelection polling organizations draw their samples from lists of currently registered voters. Such lists often provide only names and mailing addresses and no phone numbers for many people; this limits the mode of data collection to face-to-face interviewing or mailed questionnaire self-administration.

Desired Response Rate.

Face-to-face surveys routinely achieve the highest response rates, especially when conducted by the federal government. Telephone surveys typically achieve lower response rates, and Internet surveys typically achieve even lower response rates. Self-administered mail surveys can achieve high response rates if they follow an extensive protocol (Dillman, 1978; Dillman, Smyth, & Christian, 2008).

Question Form.

If a survey includes open-ended questions that require probing to clarify the details of respondents' answers, face-to-face or telephone interviewing is preferable, because interviewers can, in a standardized way, probe incomplete or ambiguous answers to ensure the usefulness and comparability of data across respondents.

Question Content.

If the issues under investigation are sensitive, self-administered questionnaires may provide respondents with a greater sense of privacy and may therefore elicit more candid responses than telephone interviews and face-to-face interviews (e.g., Bishop & Fisher, 1995; Cheng, 1988; Kreuter, Presser, & Tourangeau, 2008; Newman, Des Jarlais, Turner, Gribble, Cooley, & Paone, 2002; Wiseman, 1972).

Questionnaire Length.

Face-to-face data collection is thought to permit the longest continuous interviews – an hour or more – without respondent break-offs midway through. Telephone interviews are typically quite a bit shorter, usually lasting no more than 30 minutes, because respondents are often uncomfortable staying on the phone for longer.

Length of Data Collection Period.

Distributing questionnaires by mail requires significant amounts of time, and follow-up mailings to increase response rates further increase the overall turnaround time. Similarly, face-to-face interview surveys typically require a substantial length of time in the field. In contrast, telephone interviews and Internet surveys can be completed in very little time, within a matter of days or even hours.

Availability of Staff and Facilities.

Self-administered mail surveys require the fewest facilities and can be completed by a small staff. Face-to-face and telephone surveys typically require much larger staffs, including interviewers, their supervisors, and coordinators of the supervisors. Telephone surveys can be conducted from a central location with

sufficient office space to accommodate a staff of interviewers, but such interviewing can also be done from interviewers' homes via a central computer system that allows supervisors in a central location to monitor the interviewers' computers and conversations.

Data Quality.

A number of studies have compared the accuracy of data collected in various different modes. To date, they suggest that face-to-face interviewing may yield more representative samples and more accurate and honest reports than do telephone interviews (e.g., Holbrook, Green, & Krosnick, 2003). Computer self-administration also appears to elicit more accurate data than does telephone interviewing (e.g., Chang & Krosnick, 2009, 2010; Yeager et al., 2011). Nonetheless, telephone interviewing remains remarkably accurate in assessments of accuracy, such as predicting the outcomes of national elections.

Interviewing

When data are collected face-to-face or via telephone, interviewers play key roles. We therefore review the role of interviewers, as well as interviewer selection, training, and supervision (J. M. Converse & Schuman, 1974; Fowler & Mangione, 1986, 1990; Lavrakas, 2010; Saris, 1991).

The Role of the Interviewer.

Survey interviewers usually have three responsibilities. First, they are often responsible for locating and gaining cooperation from respondents. Second, interviewers are responsible to “train and motivate” respondents to provide thoughtful, accurate answers. Third, interviewers are responsible for executing the survey in a standardized way. The second and third responsibilities may sometimes conflict with one another. But providing explicit cues to the respondent about the requirements of the interviewing task can be done in a standardized way while still establishing rapport.

Selecting Interviewers.

It is best to use experienced, paid interviewers, rather than volunteers or students, because the former approach permits the researcher to be selective and choose only the most skilled and qualified individuals. Furthermore, volunteers or students often have an interest or stake in the substantive outcome of the

research, and they may have expectancies that can inadvertently bias data collection.

Whether they are to be paid for their work or not, all interviewers must have good reading and writing skills, and they must speak clearly. Aside from these basic requirements, few interviewer characteristics have been reliably associated with higher data quality (Bass & Tortora, 1988; Sudman & Bradburn, 1982). However, interviewer characteristics can sometimes affect answers to questions relevant to those characteristics.

One instance in which interviewer race may have had an impact along these lines involved the 1989 Virginia gubernatorial race. Preelection polls showed black candidate Douglas Wilder with a very comfortable lead over his white opponent. On election day, Wilder did win the election, but by a slim margin of 0.2%. According to Finkel, Guterbock, and Borg (1991), the overestimation of support for Wilder was attributable at least in part to social desirability. Some survey respondents apparently believed it was socially desirable to express support for the black candidate, especially when their interviewer was black. Therefore, these respondents overstated their likelihood of voting for Wilder.

Likewise, Robinson and Rohde (1946) found that the more clearly identifiable an interviewer was as being Jewish, the less likely respondents were to express anti-Jewish sentiments. Schuman and Converse (1971) found more favorable views of blacks were expressed to black interviewers, although no race-of-interviewer effects appeared on numerous items that did not explicitly ask about liking of blacks (see also Anderson, Silver, & Abramson, 1988; Cotter, Cohan, & Coulter, 1983; Davis, 1997; Davis & Silver, 2003; Hyman, Feldman, & Stember, 1954; Schaeffer, 1980). It seems impossible to eliminate the impact of interviewer race on responses, so it is preferable to randomly assign interviewers to respondents and then statistically control for interview race and the match between interviewer race and respondent race in analyses of data on race-related topics. More broadly, incorporating interviewer characteristics in statistical analyses of survey data seems well worthwhile and minimally costly.

Training Interviewers.

Interviewer training is an important predictor of data quality (Billiet & Loosveldt, 1988; Fowler & Mangione, 1986, 1990; Hansen, 2007; Lavrakas, 1993). Careful interviewer training can presumably reduce random and systematic survey error resulting from interviewer mistakes and nonstandardized survey implementation across interviewers. It seems worth the effort, then, to

conduct thorough, well-designed training sessions, especially when one is using inexperienced and unpaid interviewers (e.g., students as part of a class project). Training programs last two days or longer at some survey research organizations, because shorter training programs do not adequately prepare interviewers, resulting in substantial reductions in data quality (Fowler & Mangione, 1986, 1990).

In almost all cases, training should cover topics such as

1. how to use all interviewing equipment;
2. procedures for randomly selecting respondents within households;
3. techniques for eliciting survey participation and avoiding refusals;
4. opportunities to gain familiarity with the survey instrument and to practice administering the questionnaire;
5. instructions regarding how and when to probe incomplete responses;
6. instructions on how to record answers to open-and closed-ended questions; and
7. guidelines for establishing rapport while maintaining a standardized interviewing atmosphere.

Training procedures can take many forms (e.g., lectures, written training materials, observation of real or simulated interviews). It is important that the training session involve supervised practice interviewing. Pairs of trainees are routinely asked to take turns playing the roles of interviewer and respondent. Such role playing might also involve the use of various “respondent scripts” that present potential problems for the interviewer to practice handling. Interviewers can be trained both in how to recruit potential respondents and in how to ask the survey's questions.

Supervision.

Carefully monitoring ongoing data collection permits early detection of problems and seems likely to improve data quality. In face-to-face or telephone surveys, researchers should maintain running estimates of each interviewer's average response rate, level of productivity, and cost per completed interview, to identify potential problems. Researchers can also monitor the data collected by interviewers in real time, to be sure that open-ended answers are being transcribed properly, for example (cf., Steve et al., 2008).

The quality of each interviewer's completed questionnaires should be

monitored, and if possible, some of the interviews themselves should be supervised. When surveys are conducted by telephone, monitoring the interviews is relatively easy and inexpensive and should be done routinely. When interviews are conducted face-to-face, interviewers can make audio recordings of some or all of their interviews to permit evaluation of each aspect of the interview.

Validation.

When data collection occurs from a single location (e.g., telephone interviews that are conducted from a central phone bank), researchers can be relatively certain that the data are authentic. When data collection does not occur from a central location (e.g., face-to-face interviews or telephone interviews conducted from interviewers' homes), researchers might be less certain. It may be tempting for some interviewers to falsify some of the questionnaires that they turn in, and some occasionally do. This is referred to as curbstoning, a topic addressed in a 2009 report by the American Association for Public Opinion Research, the nation's leading professional association of survey researchers (see <http://www.aapor.org/Content/aapor/AdvocacyandInitiatives/StandardsandEthics>). To guard against this, researchers often establish a procedure for confirming that a randomly selected subset of all interviews did indeed occur (e.g., recontacting some respondents and asking them about whether the interview took place and how long it lasted). This can only be accomplished if contact information for respondents is maintained by the researcher, which must be done carefully in order to keep identities confidential and never connected to responses.

Total Survey Error

As is no doubt obvious by now, high-quality survey data collection can be very costly. And many survey researchers, even those with big budgets, nonetheless have limits on how much they can spend on a project. Such financial limitations routinely force researchers to make choices about how to spend their money. That is, a decision to spend money on one component of a study (e.g., paying financial incentives) will necessarily mean that less money is available to spend on other aspects of the data collection effort. For example, in order to be able to pay for an extra day of interviewer training, a researcher might have to reduce the number of interviewer hours available for conducting the survey's interviews.

How should a researcher go about making these choices? What principles

should guide the allocation of resources? Building on the work of Hansen (e.g., Hansen & Madow, 1953), the “total survey error” perspective suggests that such decisions should be made in ways that maximize the accuracy of the obtained data (cf. Dillman, 1978, Fowler, 1988; Groves, 1989). The total survey error perspective is based on the notion that the ultimate goal of survey research is to accurately measure particular constructs within a sample of people who represent the population of interest. In any given survey, the overall deviation from this ideal is the cumulative result of several sources of survey error.

Specifically, the total survey error perspective disaggregates overall error into seven major components: coverage error, sampling error, nonresponse error, specification error, measurement error, adjustment error, and processing error. *Coverage error* refers to the bias that can result when the pool of potential respondents from which a sample is selected does not include some portions of the population of interest. *Sampling error* refers to the random differences that invariably exist between any sample and the population from which it was selected. *Nonresponse error* is the bias that can result when data are not collected from all members of a sample. *Specification error* refers to how well the constructs the researchers purport to have assessed were actually measures. And *measurement error* refers to all distortions in the assessment of the construct of interest, including systematic biases and random variance that can be brought about by respondents’ own behavior (e.g., misreporting true attitudes, failing to pay close attention to a question), interviewer behavior (e.g., misrecording responses, providing cues that lead respondents to respond in one way or another), and the questionnaire (e.g., ambiguous or confusing wording, biased question wording or response options). *Adjustment error* refers to the statistical corrections that have been made to address issues such as unequal probabilities of selection at the time of sampling, and the problems that noncoverage and nonresponse may have caused in yielding a final sample that is not representative of the target population. *Processing error* refers to how well the “raw dataset” has been cleaned and processed (e.g., the coding of open-ended verbatim transcripts, the transformation of data gathered at the interval level into categorical variables, the creation of multi-item scales, etc.) in creating a final data set that researchers will analyze.

The total survey error perspective advocates explicitly taking into consideration each of these sources of error and making decisions about the allocation of finite resources with the goal of reducing the sum of the seven. There is remarkably little scientific evidence available at the moment to quantify the gains in data accuracy that result from a dollar being spent in various

different ways, so researchers are currently left to make guesses about the benefits of various types of expenditures. Perhaps in the future, more empirical research will be done to guide these sorts of decisions.

Conclusions

This chapter offers only the very beginning of an introduction to the process of survey data collection, and interested readers can turn to more extensive treatments of these issues to gain further mastery (e.g., Groves, Fowler, Couper, Lepkowski, Singer, & Tourangeau, 2009; Lavrakas, 2008). We hope here to have illustrated for social psychologists just how complex and challenging the survey data collection process is and why it is worth the trouble. Social science, after all, is meant to gain insights into populations of people, and the organizations that fund our work deserve to know whether our findings accurately describe entire populations or describe only narrow subsets. Given the strong traditions of representative sampling in other disciplines, such as political science, sociology, and economics, psychologists place themselves at a disadvantage if they completely ignore the scientific imperative to confirm the generalizability and applicability of their findings in rigorous ways. Surveys offer the opportunity to do just this.

But even for a social-personality psychologist who chooses to forego studies of representative samples, the survey methodology literature has a lot to offer to help that person do his or her work more effectively. Specifically, the huge and growing literature on questionnaire design offers guidelines for optimizing measurement in laboratory experiments of convenience samples. To ignore that literature is to risk using measuring tools that acquire a great deal of random or nonrandom measurement error and therefore make it more difficult for a researcher to detect real relationships between variables. Therefore, in the interest of minimizing the number of respondents in a study while maximizing the ability to gauge effect sizes accurately, questions should be designed according to the principles evolving in the survey questionnaire design arena. Just one striking example of suboptimality is social-personality psychologists' continued reliance on agree-disagree rating scales in the face of the huge literature documenting the damaging impact of acquiescence response bias. We look forward to improved measurement by social psychologists and enhanced efficiency of the scientific inquiry that is likely to result, and such changes can be spurred by careful attention to the findings of survey methodologists on questionnaire design.

Furthermore, it is beyond dispute that the diversity of life experiences, perspectives, and approaches to decision making is much, much greater in the entire adult population than in the subpopulation of students enrolled in college psychology courses. Therefore, surveys of general population samples offer not only the opportunity to produce findings legitimating generalization, but also exciting opportunities for theory development. By collecting and analyzing data from large, heterogeneous full-population samples, researchers can be spurred to consider a wide range of new moderator variables that may encapsulate the impact of life situations and individual attributes on social psychological processes. And as a result, our theories may end up being richer, especially thanks to the emergence of new technologies that permit collection of many sorts of data in the course of daily life from representative samples of people.

But even if occasional use of survey data from representative samples does not change the nature of our scientific findings at all, it seems likely that use of such data will (appropriately) enhance the perceived credibility of our enterprise and will illustrate that social psychologists are willing to invest the effort and funds necessary to move beyond convenient laboratory studies of captive audiences in order to objectively evaluate the applicability of our claims. This, in and of itself, seems like sufficient justification for social psychologists to learn about how survey data are collected, learn about how to analyze them properly, make funding requests that are sufficient to allow such work, and enrich our work product as a result.

References

- Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. (1950). *The authoritarian personality*. New York: Harper & Row.
- Ahlgren, A. (1983). Sex differences in the correlates of cooperative and competitive school attitudes. *Developmental Psychology*, 19, 881–888.
- Alderman, H., Behrman, J. R., Kohler, H., Maluccio, J. A., & Watkins, S. C. (2001). Attrition in longitudinal household survey data. *Demographic Research*, 5(4), 79–124.
- Alwin, D. F., Cohen, R. L., & Newcomb, T. M. (1991). *The women of Bennington: A study of political orientations over the life span*. Madison: University of Wisconsin Press.
- Alwin, D. F., & Jackson, D. J. (1982). Adult values for children: An application

- of factor analysis to ranked preference data. In R. M. Hauser, D. Mechanic, A. O. Haller, & T. S. Hauser (Eds.), *Sociological structure and behavior: Essays in honor of William Hamilton Sewell* (pp. 311–329). New York: Academic Press.
- Alwin, D. F., & Krosnick, (1991). The reliability of survey attitude measurement. The influence of question and respondent attributes. *Sociological Methods and Research*, 20, 139–181.
- Anderson, B., Silver, B., & Abramson, P. (1988). The effects of the race of the interviewer on measures of electoral participation by blacks in SRC national election studies. *Public Opinion Quarterly*, 52, 53–83.
- Babbie, E. R. (1990). *Survey research methods*. Belmont, CA: Wadsworth.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Bass, R. T., & Tortora, R. D. (1988). A comparison of centralized CATI facilities for an agricultural labor survey. In R. M. Groves, P. P. Beimer, L. E. Lyberg, J. T. Massey, W. L. Nicholls, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 497–508). New York: Wiley.
- Bearden, W. Q., Netemeyer, R. G., & Mobley, M. F. (1993). *Handbook of marketing scales*. Newbury Park, CA: Sage.
- Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71, 287–311.
- Beckett, S., Gould, W., Lillard, L., & Welch, F. (1988). The PSID after fourteen years: An evaluation. *Journal of Labor Economics*, 6(4), 472–492.
- Bem, D. J., & McConnell, H. K. (1970.). Testing the self-perception explanation of dissonance phenomena: On the salience of premanipulation attitudes. *Journal of Personality and Social Psychology*, 14, 23–31.
- Benet, V., & Waller, N. G. (1995). The big seven factor model of personality description: Evidence for its cross-cultural generality in a Spanish sample. *Journal of Personality and Social Psychology*, 69, 701–718.
- Berinsky, A. J. (1999). The two faces of public opinion. *American Journal of*

Political Science, 43, 1209–1230.

- Billiet, J., & Loosveldt, G. (1988). Improvement of the quality of responses to factual survey questions by interviewer training. *Public Opinion Quarterly*, 52, 190–211.
- Bischoping, K. (1989). An evaluation of interviewer debriefing in survey pretests. In C. F. Cannell, L. Oskenberg, F. J. Fowler, G. Kalton, & K. Bischoping (Eds.), *New techniques for pretesting survey questions* (pp. 15–29). Ann Arbor, MI: Survey Research Center.
- Bishop, G. F., & Fisher, B. S. (1995). “Secret ballots” and self-reports in an exit-poll experiment. *Public Opinion Quarterly*, 59, 568–588.
- Blalock, H. M. (1972). *Causal inferences in nonexperimental research*. New York: Norton.
- Blalock, H. M. (1985). *Causal models in panel and experimental designs*. New York: Aldine.
- Blumberg, S. J., & Luke, J. V. (2012). *Wireless substitution: Early release of estimates from the National Health Interview Survey, January–June 2012*. Atlanta, GA: Centers for Disease Control and Prevention.
- Bogaert, A. F. (1996). Volunteer bias in human sexuality research: Evidence for both sexuality and personality differences in males. *Archives of Sexual Behavior*, 25, 125–140.
- Box-Steffensmeier, J. M., Jacobson, G. C., & Grant, J. T. (2000). Question wording and the house vote choice: Some experimental evidence. *Public Opinion Quarterly*, 64, 257–270.
- Bradburn, N. M., & Sudman, S. (1979). *Improving interview method and questionnaire design*. San Francisco: Jossey-Bass.
- Bradburn, N. M., Sudman, S., & Associates. (1981). *Improving interview method and questionnaire design*. San Francisco: Jossey-Bass.
- Brehm, J. (1993). *The phantom respondents*. Ann Arbor: University of Michigan Press.
- Brehm, J., & Rahn, W. (1997). Individual-level evidence for the causes and consequences of social capital. *American Journal of Political Science*, 41, 999–1023.

- Bridge, R. G., Reeder, L. G., Kanouse, D., Kinder, D. R., Nagy, V. T., & Judd, C. M. (1977). Interviewing changes attitudes – sometimes. *Public Opinion Quarterly*, 41, 56–64.
- Byrne, D. (1971). *The attraction paradigm*. New York: Academic Press.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and divergent validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.
- Cannell, C. F., Miller, P., & Oskenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 389–437). San Francisco: Jossey-Bass.
- Cannell, C. F., Oksenberg, L., & Converse, J. M. (1977). *Experiments in interviewing techniques: Field experiments in health reporting, 1971–1977*. Hyattsville, MD: National Center for Health Services Research.
- Cantor, D., O'Hare, B., & O'Connor, K. (2008). The use of monetary incentives to reduce nonresponse in random digit dial telephone surveys. In J. M. Lepkowski *et al.* (Eds.), *Advances in telephone survey methodology* (pp. 471–498). New York: Wiley.
- Caspi, A., Bem, D. J., & Elder, G. H., Jr. (1989). Continuities and consequences of interactional styles across the life course. *Journal of Personality*, 57, 375–406.
- Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing vs. the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, 73, 641–678.
- Chang, L., & Krosnick, J. A. (2010). Comparing oral interviewing with self-administered computerized questionnaires: An experiment. *Public Opinion Quarterly*, 74, 154–167.
- Chanley, V. A., Rudolph, T. J., & Rahn, W. M. (2000). The origins and consequences of public trust in government. A time series analysis. *Public Opinion Quarterly*, 64, 239–256.
- Cheng, S. (1988). Subjective quality of life in the planning and evaluation of

- programs. *Evaluation and Program Planning*, 11, 123–134.
- Clinton, J. D. (2001). *Panel bias from attrition and conditioning: A case study of the knowledge networks panel*. Stanford, CA: Stanford University Press.
- Clinton, J. D., & Rogers, S. (2012). *Robo-polls: Taking cues from traditional sources?* Unpublished manuscript, Vanderbilt University, Nashville, TN.
- Cohen, D., Nisbett, R. E., Bowdle, B. F., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honor: An “experimental ethnography.” *Journal of Personality and Social Psychology*, 70, 945–960.
- Congressional Information Service. (1990). *American statistical index*. Bethesda, MD: Author.
- Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Beverly Hills, CA: Sage.
- Converse, J. M., & Schuman, H. (1974). *Conversations at random*. New York: Wiley.
- Converse, P. E. (1964). The nature of belief systems in the mass public. In D. E. Apter (Ed.), *Ideology and discontent* (pp. 206–261). New York: Free Press.
- Cook, A. R., & Campbell, D. T. (1969). *Quasi-experiments: Design and analysis issues for field settings*. Skokie, IL: Rand McNally.
- Cook, C., Heath, F., & Thompson, R. L. (2001). Score reliability in web-or Internet-based surveys: Unnumbered graphic rating scales versus Likert-type scales. *Educational and Psychological Measurement*, 61, 697–706.
- Coombs, C. H., & Coombs, L. C. (1976). “Don't know”: Item ambiguity or respondent uncertainty? *Public Opinion Quarterly*, 40, 497–514.
- Cooper, D. R., & Clare, D. A. (1981). A magnitude estimation scale for human values. *Psychological Reports*, 49, 431–438.
- Costa, P. T., McCrae, R. R., & Arenberg, D. (1983). Recent longitudinal research on personality and aging. In K. W. Schaie (Ed.), *Longitudinal studies of adult psychological development* (pp. 222–263). New York: Guilford Press.
- Cotter, P., Cohen, J. & Coulter, P. (1982). Race-of-interviewer effects in telephone interviews. *Public Opinion Quarterly*, 46, 278–284.
- Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response

- set as a personality variable. *Journal of Abnormal and Social Psychology*, 60, 151–174.
- Couper, M. (2008). *Designing effective web surveys*. Cambridge: Cambridge University Press.
- Couper, M. P., Tourangeau, R., Conrad, R. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review*, 24, 227–245.
- Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly*, 65, 230–253.
- Courneya, K. S., Jones, L. W., Rhodes, R. E., & Blanchard, C. M. (2003). Effect of response scales on self-reported exercise frequency. *American Journal of Health Behavior*, 27, 613–622.
- Coye, R. W. (1985). Characteristics of participants and nonparticipants in experimental research. *Psychological Reports*, 56, 19–25.
- Crano, W. D., & Brewer, M. B. (1986). *Principals and methods of social research*. Newton, MA: Allyn and Bacon.
- Curtin, R., Presser, S., & Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, 64, 413–428.
- Davies, T., & Gangadharan, S. P. (Eds.). (2009). *Online deliberation: Design, research, and practice*. Stanford, CA: CSLI Publications.
- Davis, D. W. (1997). Nonrandom measurement error and race of interviewer effects among African Americans. *Public Opinion Quarterly*, 61, 183–207.
- Davis, D. W., & Silver, B. D. (2003). Stereotype threat and race of interviewer effects in a survey on political knowledge. *American Journal of Political Science*, 47, 33–45.
- DeBell, M., & Krosnick, J. A. (2009). *Computing weights for American National Election Study survey data*. ANES Technical Report series, no. nes012427. Ann Arbor, MI, and Palo Alto, CA: American National Election Studies. Retrieved August 26, 2013, from <http://www.electionstudies.org/resources/papers/nes012427.pdf>.
- Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating and mixed standard scale formats. *Journal of Applied*

Psychology, 65, 147–154.

Dillman, D., Smyth, J.D., & Christian, L. M. (2008). *Internet, mail, and mixed-mode surveys: The tailored design method*. New York: Wiley.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.

Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design method* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Dollinger, S. J., & Leong, F. T. (1993). Volunteer bias and the five-factor model. *Journal of Psychology*, 127, 29–36.

Donovan, R. J., & Leivers, S. (1993). Using paid advertising to modify racial stereotype beliefs. *Public Opinion Quarterly*, 57, 205–218.

Ebel, R. L. (1982). Proposed solutions to two problems of test construction. *Journal of Educational Measurement*, 19, 267–278.

Ehrlich, H. J. (1964). Instrument error and the study of prejudice. *Social Forces*, 43, 197–206.

Eifermann, R. R. (1961). Negation: A linguistic variable. *Acta Psychologica*, 18, 258–273.

Elig, T. W., & Frieze, I. H. (1979). Measuring causal attributions for success and failure. *Journal of Personality and Social Psychology*, 37, 621–634.

England, L. R. (1948). Capital punishment and open-end questions. *Public Opinion Quarterly*, 12, 412–416.

Eveland, Jr., W. P., Hayes, A. F., Shah, D. V., & Kwak, N. (2005). Understanding the relationship between communication and political knowledge: A model comparison approach using panel data. *Political Communication*, 22, 423–446.

Falaris, E. M., & Peters, H. E. (1998). Survey attrition and schooling choices. *The Journal of Human Resources*, 33, 531–554.

Finkel, S. E., Guterbock, T. M., & Borg, M. J. (1991). Race-of-interviewer effects in a preelection poll: Virginia 1989. *Public Opinion Quarterly*, 55, 313–330.

Fitzgerald, J., Gottschalk, P., & Moffitt, R. (1998a). An analysis of sample

attrition in panel data: The Michigan panel study of income dynamics. *NBER Technical Working Papers*, National Bureau of Economic Research, Inc.

Fitzgerald, J., Gottschalk, P., & Moffitt, R. (1998b). An analysis of the impact of sample attrition on the second generation of respondents in the Michigan panel study of income dynamics. *The Journal of Human Resources*, 33, 300–344.

Forsyth, B. H., & Lessler, J. T. (1991). Cognitive laboratory methods: A taxonomy. In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement error in surveys* (pp. 393–418). New York: Wiley.

Fowler, F. J. (1988). *Survey research methods* (2nd ed.). Beverly Hills, CA: Sage.

Fowler, F. J. (2009). *Survey research methods* (4th ed.). Thousand Oaks, CA: Sage.

Fowler, Jr., F. J., & Mangione, T. W. (1986). *Reducing interviewer effects on health survey data*. Washington, DC: National Center for Health Statistics.

Fowler, Jr., F. J., & Mangione, T. W. (1990). *Standardized survey interviewing*. Newbury Park, CA: Sage.

Frey, J. H. (1989). *Survey research by telephone* (2nd ed.). Newbury Park, CA: Sage.

Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, 72, 892–913.

Geer, J. G. (1988). What do open-ended questions measure? *Public Opinion Quarterly*, 52, 365–371.

Glenn, N. O. (1980). Values, attitudes, and beliefs. In O. G. Brim & J. Kagan (Eds.), *Constancy and change in human development* (pp. 596–640). Cambridge, MA: Harvard University Press.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96, 1029–1046.

Granberg, D. (1985). An anomaly in political perception. *Public Opinion Quarterly*, 49, 504–516.

- Granberg, D., & Holmberg, S. (1992). The Hawthorne effect in election studies: The impact of survey participation on voting. *British Journal of Political Science*, 22, 240–247.
- Green, D. P., & Gerber, A. S. (2006). Can registration-based sampling improve the accuracy of midterm election forecasts? *Public Opinion Quarterly*, 70, 197–223.
- Greenwald, A. G., Carnot, C. G., Beach, R., & Young, B. (1987). Increasing voting behavior by asking people if they expect to vote. *Journal of Applied Psychology*, 72, 315–318.
- Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.
- Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70, 646–675.
- Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, 56, 475–495.
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. New York: Wiley.
- Groves, R. M., & Kahn, R. L. (1979). *Surveys by telephone: A national comparison with personal interviews*. New York: Wiley.
- Haddock, G., & Zanna, M. P. (1998). On the use of open-ended measures to assess attitudinal components. *British Journal of Social Psychology*, 37, 129–149.
- Hamamura, T., Meijer, Z., Heine, S. J., Kamaya, K., & Hori, I. (2009). Approach-avoidance motivation and information processing: A cross-cultural analysis. *Personality and Social Psychology Bulletin*, 35, 454–462.
- Han, S., & Shavitt, S. (1994). Persuasion and culture: Advertising appeals in individualistic and collectivist societies. *Journal of Experimental and Social Psychology*, 30, 326–350.
- Hansen, K. M. (2007). The effects of incentives, interview length, and interviewer characteristics on response rates in CATI-study. *International Journal of Public Opinion Research*, 19, 112–121.
- Hansen, M. H., & Madow, W. G. (1953). *Survey methods and theory*. New York: Wiley.

- Harzing, A.-W., Baldueza, J., Barner-Rasmussen, W., Barzantny, C., Canabal, A. *et al.* (2009). Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International Business Review*, 18, 417–432.
- Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44 (2), 174–199.
- Heckathorn, D. D. (2002). Respondent-driven sampling II: Deriving valid estimates from chain-referral samples of hidden populations. *Social Problems*, 49(1), 11–34.
- Heine, S. J., & Lehman, D. R. (1995). Cultural variation in unrealistic optimism: Does the west feel more invulnerable than the east? *Journal of Personality and Social Psychology*, 68, 595–607.
- Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.
- Hess, J., Singer, E., & Bushery, J. (1999). Predicting test-retest reliability from behavior coding. *International Journal of Public Opinion Research*, 11, 346–360.
- Himmelfarb, S., & Norris, F. H. (1987). An examination of testing effects in a panel study of older persons. *Personality and Social Psychology Bulletin*, 13, 188–209.
- Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone vs. face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, 67, 79–125.
- Holbrook A. L., Krosnick, J. A., Moore, D., & Tourangeau, R. (2007). Response order effects in dichotomous categorical questions presented orally: The impact of questiona and respondent attributes. *Public Opinion Quarterly*, 71, 325–348.
- Holbrook, A. L., Krosnick, J. A., & Pfent, A. M. (2008). The causes and consequences of response rates in surveys by the news media and government contractor survey research firms. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. De Leeuw, L. Japac, P. J. Lavrakas, M. W. Link, & R. L. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 499–528). New York: Wiley.

- Hothersall, D. (1984). *History of psychology*. New York: Random House.
- Hovland, C. I., Harvey, O. J., & Sherif, M. (1957). Assimilation and contrast effects in reactions to communication and attitude change. *Journal of Personality and Social Psychology*, 55, 244–252.
- Hurd, A.W. (1932). Comparisons of short answer and multiple choice tests covering identical subject content. *Journal of Educational Psychology*, 26, 28–30.
- Hurd, M. D. (1999). Anchoring and acquiescence bias in measuring assets in household surveys. *Journal of Risk and Uncertainty*, 19, 111–136.
- Hyman, H. A., Feldman, J., & Stember, C. (1954). *Interviewing in social research*. Chicago: University of Chicago Press.
- Jackman, M. R. (1973). Education and prejudice or education and response-set? *American Sociological Review*, 38, 327–339.
- Jackson, J. E. (1979). Bias in closed-ended issue questions. *Political Methodology*, 6, 393–424.
- James, L. R., & Singh, B. H. (1978). An introduction to the logic, assumptions, and the basic analytic procedures of two-stage least squares. *Psychological Bulletin*, 85, 1104–1122.
- Jenkins, J. G. (1935). *Psychology in business and industry*. New York: Wiley.
- Judd, C. M., & Johnson, J. T. (1981). Attitudes, polarization, and diagnosticity: Exploring the effect of affect. *Journal of Personality and Social Psychology*, 41, 26–36.
- Kalton, G. (1983). *Introduction to survey sampling*. Beverly Hills, CA: Sage.
- Kam, C. D., & Ramos, J. M. (2008). Joining and leaving the rally. *Public Opinion Quarterly*, 72, 619–650.
- Katz, D. (1942). Do interviewers bias poll results? *Public Opinion Quarterly*, 6, 248–268.
- Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S., 2000; Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64, 125–148.

- Kellstedt, P. M., Peterson, D. A. M., & Ramirez, M. D. (2010). The macro politics of gender gap. *Public Opinion Quarterly*, 74, 477–498.
- Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley.
- Kessler, R. C., & Greenberg, D. F. (1981). *Linear panel analysis: Models of quantitative change*. New York: Academic Press.
- Kim, S.-H., Scheufele, D. A., Shanahan, J., & Choi, D.-H. (2011). Deliberation in spite of controversy? News media and the public's evaluation of a controversial issue in South Korea. *Journalism & Mass Communication Quarterly*, 88, 2320–2336.
- Kinder, D. R. (1978). Political person perception: The asymmetrical influence of sentiment and choice on perceptions of presidential candidates. *Journal of Personality and Social Psychology*, 36, 859–871.
- Kinder, D. R., & Sanders, L. M. (1990). Mimicking political debate within survey questions: The case of White opinion on affirmative action for Blacks. *Social Cognition*, 8, 73–103.
- Kish, L. (1965). *Survey sampling*. New York: Wiley.
- Kitayama, S., & Markus, H. R. (1994). *Emotion and culture: Empirical studies of mutual influence*. Washington, DC: American Psychological Association.
- Kitayama, S., Park, H., Sevincer, A. T., Karasawa, M., & Uskul, A. K. (2009). A cultural task analysis of implicit independence: Comparing North America, Western Europe, and East Asia. *Journal of Personality and Social Psychology*, 97, 236–255.
- Klockars, A. J., & Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement*, 25, 85–96.
- Kraut, R. E., & McConahay, J. B. (1973). How being interviewed affects voting: An experiment. *Public Opinion Quarterly*, 37, 398–406.
- Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, 72, 847–865.
- Krosnick, J. A. (1988a). Attitude importance and attitude change. *Journal of Experimental Social Psychology*, 24, 240–255.

- Krosnick, J. A. (1988b). The role of attitude importance in social evaluation: A study of policy preferences, presidential candidate evaluations, and voting behavior. *Journal of Personality and Social Psychology*, 55, 196–210.
- Krosnick, J. A. (1991a). Americans' perceptions of presidential candidates: A test of the projection hypothesis. *Journal of Social Issues*, 46, 159–182.
- Krosnick, J. A. (1991b). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213–236.
- Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly*, 51, 201–219.
- Krosnick, J. A., & Alwin, D. F. (1988). A test of the form resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, 52, 526–538.
- Krosnick, J. A., & Alwin, D. F. (1989). Aging and susceptibility to attitude change. *Journal of Personality and Social Psychology*, 57, 416–425.
- Krosnick, J. A., & Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science*, 37, 941–964.
- Krosnick, J. A., Boninger, D. S., Chuang, Y. C., Berent, M. K., & Carnot, C. G. (1993). Attitude strength: One construct or many related constructs? *Journal of Personality and Social Psychology*, 65, 1132–1151.
- Krosnick, J. A., & Brannon, L. A. (1993). The impact of the Gulf War on the ingredients of presidential evaluations: Multidimensional effects of political involvement. *American Political Science Review*, 87, 963–975.
- Krosnick, J. A., & Fabrigar, L. R. (forthcoming). *Designing great questionnaires: Insights from psychology*. New York: Oxford University Press.
- Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, R. C., Presser, S., Ruud, P. A., Smith, V. K., Moody, W. R., Green, M. C., & Conaway, M. (2002). The impact of “no opinion” response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, 66, 371–403.

- Krosnick, J. A., & Kinder, D. R. (1990). Altering popular support for the president through priming: The Iran-Contra affair. *American Political Science Review*, 84, 497–512.
- Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An overview. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 1–24). Hillsdale, NJ: Erlbaum.
- Laumann, E. O., Michael, R. T., Gagnon, J. H., & Michaels, S. (1994). *The social organization of sexuality: Sexual practices in the United States*. Chicago: University of Chicago Press.
- Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.
- Lavrakas, P. J. (Ed.). (2008). *Encyclopedia of survey research methods*. Thousand Oaks, CA: Sage.
- Lavrakas, P. J. (2010). Telephone surveys. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research*. San Diego, CA: Elsevier.
- Li, X. (2008). Third-person effect, optimistic bias, and sufficiency resource in Internet use. *Journal of Communication*, 58, 568–587.
- Lin, I. F., & Shaeffer, N. C. (1995). Using survey participants to estimate the impact of nonparticipation. *Public Opinion Quarterly*, 59, 236–258.
- Lindzey, G. E., & Guest, L. (1951). To repeat – checklists can be dangerous. *Public Opinion Quarterly*, 15, 355–358.
- Link, M. W., Battaglia, M. P., Frankel, M. R., Osborn, L., & Mokdad, A. H. (2007). Reaching the U.S. cell phone generation: Comparison of cell phone survey results with an ongoing landline telephone survey. *Public Opinion Quarterly*, 71, 814–839.
- Lodge, M. (1981). *Magnitude scaling: Quantitative measurement of opinions*. Beverly Hills, CA: Sage.
- Lozano, L. M., Garcia-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, 4(2), 73–79.
- Luskin, R. C., & Bullock, J. G. (2011). “Don't know” means “don't know”: DK responses and the public's level of political knowledge. *Journal of Politics*, 73,

547–557.

- Mann, C. B. (2005). Unintentional voter mobilization: Does participation in preelection surveys increase voter turnout? *The Annals of the American Academy of Political and Social Science*, 601, 155–168.
- Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Journal of Personality and Social Psychology*, 98, 224–253.
- Marquis, K. H., Cannell, C. F., & Laurent, A. (1972). Reporting for health events in household interviews: Effects of reinforcement, question length, and reinterviews. In *Vital and health statistics* (Series 2, No. 45) (pp. 1–70). Washington, DC: U.S. Government Printing Office.
- Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert Scale items? Study I: Reliability and validity. *Educational and Psychological Measurement*, 31, 657–674.
- McClendon, M. J. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods and Research*, 20, 60–103.
- McClendon, M. J., & Alwin, D. F. (1993). No-opinion filters and attitude measurement reliability. *Sociological Methods and Research*, 21, 438–464.
- McFarland, S. G. (1981). Effects of question order on survey responses. *Public Opinion Quarterly*, 45, 208–215.
- Merkle, D., & Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.). *Survey Nonresponse* (pp. 243–258). New York: Wiley.
- Miethe, T. D. (1985). The validity and reliability of value measurements. *Journal of Personality*, 119, 441–453.
- Miller, J. M., & Krosnick, J. A. (1998). The impact of candidate name order on election outcomes. *Public Opinion Quarterly*, 62, 291–330.
- Mirowsky, J., & Ross, C. E. (1991). Eliminating defense and agreement bias from measures of the sense of control: A 2 × 2 index. *Social Psychology Quarterly*, 54, 127–145.
- Moore, D. W. (2002). New types of question-order effects: Additive and subtractive. *Public Opinion Quarterly*, 66, 80–91.

- Mortimer, J. T., Finch, M. D., & Kumka, D. (1982). Persistence and change in development: The multidimensional self-concept. In P. B. Baltes & O. G. Brim, Jr. (Eds.), *Lifespan development and behavior* (Vol. 4, pp. 263–312). New York: Academic Press.
- Mosteller, F., Hyman, H., McCarthy, P. J., Marks, E. S., & Truman, D. B. (1949). *The preelection polls of 1948: Report to the committee on analysis of preelection polls and forecasts*. New York: Social Science Research Council.
- Munson, J. M., & McIntyre, S. H. (1979). Developing practical procedures for the measurement of personal values in cross-cultural marketing. *Journal of Marketing Research*, 16, 48–52.
- Myers, J. H., & Warner, W. G. (1968). Semantic properties of selected evaluation adjectives. *Journal of Marketing Research*, 5, 409–412.
- Nathan, B. R., & Alexander, R. A. (1985). The role of inferential accuracy in performance rating. *Academy of Management Review*, 10, 109–115.
- Nelson, D. (1985). Informal testing as a means of questionnaire development. *Journal of Official Statistics*, 1, 179–188.
- Nesselroade, J. R., & Baltes, P. B. (1974). Adolescent personality development and historical change: 1970–1972. *Monographs of the Society for Research in Child Development*, 39 (No. 1, Serial No. 154).
- Newman, J. C., Des Jarlais, D. C., Turner, C. F., Gribble, J., Cooley, P., & Paone, D. (2002). The differential effects of face-to-face and computer interview modes. *American Journal of Public Health*, 92, 294–297.
- Nisbett, R. E. (1993). Violence and U.S. regional culture. *American Psychologist*, 48, 441–449.
- Nisbett, R. E., & Cohen, D. (1996). *Culture of honor: The psychology of violence in the south*. Boulder, CO: Westview Press.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychology Review*, 84, 231–259.
- Office of Information and Regulatory Affairs. (2006). Questions and answers when designing surveys for information collections. Washington, DC: Office of Management and Budget. Retrieved August 26, 2013, from http://www.whitehouse.gov/sites/default/files/omb/inforeg/pmc_survey_guida

- Pasek, J. (2012a). Package “anesrake.” Retrieved August 26, 2013, from <http://cran.r-project.org/web/packages/anesrake/anesrake.pdf>.
- Pasek, J. (2012b). Online weighting tool. Retrieved August 26, 2013, from <http://joshpasek.com/category/software/>
- Pasek, J., Tahk, A., Lelkes, Y., Krosnick, J. A., Payne, K., Akhtar, O., & Tompson, T. (2009). Determinants of turnout and candidate choice in the 2008 U.S. Presidential election: Illuminating the impact of racial prejudice and other considerations. *Public Opinion Quarterly*, 73, 943–994.
- Patrick, D. L., Bush, J. W., & Chen, M. M. (1973). Methods for measuring levels of well-being for a health status index. *Health Services Research*, 8, 228–245.
- Payne, B. K., Krosnick, J. A., Pasek, J., Lelkes, Y., Akhtar, O., & Tompson, T. (2010). Implicit and explicit prejudice in the 2008 American presidential election. *Journal of Experimental Social Psychology*, 46, 367–374.
- Payne, S. L. (1949/1950). Case study in question complexity. *Public Opinion Quarterly*, 13, 653–658.
- Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag.
- Petty, R. E., & Cacioppo, J. T. (1996). Addressing disturbing and disturbed consumer behavior: Is it necessary to change the way we conduct behavioral science? *Journal of Marketing Research*, 33, 1–8.
- Presser, S. (1990). Measurement issues in the study of social change. *Social Forces*, 68, 856–868.
- Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results? In P. V. Marsden (Ed.), *Sociological methodology, 1994* (pp. 73–104). Cambridge, MA: Blackwell.
- Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., & Singer, E. (2004). *Methods for testing and evaluating survey questionnaires*. New York: Wiley-Interscience.
- Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104, 1–15.

- Rahn, W. M., Krosnick, J. A., & Breuning, M. (1994). Rationalization and derivation processes in survey studies of political candidate evaluation. *American Journal of Political Science*, 38, 582–600.
- Rankin, W. L., & Grube, J. W. (1980). A comparison of ranking and rating procedures for value system measurement. *European Journal of Social Psychology*, 10, 233–246.
- Remmers, H. H., Marschat, L. E., Brown, A., & Chapman, I. (1923). An experimental study of the relative difficulty of true-false, multiple-choice, and incomplete-sentence types of examination questions. *Journal of Educational Psychology*, 14, 367–372.
- Reynolds, T. J., & Jolly, J. P. (1980). Measuring personal values: An evaluation of alternative methods. *Journal of Marketing Research*, 17, 531–536.
- Rhee, E., Uleman, J. S., Lee, H. K., & Roman, R. J. (1995). Spontaneous self-descriptions and ethnic identities in individualistic and collectivist cultures. *Journal of Personality and Social Psychology*, 69, 142–152.
- Richardson, J. D. (2004). Isolating frequency scale effects on self-reported loneliness. *Personality and Individual Differences*, 36, 235–244.
- Roberto, K. A., & Scott, J. P. (1986). Confronting widowhood: The influence of informal supports. *American Behavioral Scientist*, 29, 497–511.
- Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: construct validation of a single-item measure and the Rosenberg self-esteem scale. *Personality and Social Psychology Bulletin*, 27, 151–161.
- Robinson, D., & Rohde, S. (1946). Two experiments with an anti-semitism poll. *Journal of Abnormal and Social Psychology*, 41, 136–144.
- Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review*, 96, 341–357.
- Ross, M. W. (1988). Prevalence of classes of risk behaviors for HIV infection in a randomly selected Australian population. *Journal of Sex Research*, 25, 441–450.
- Ruch, G. M., & DeGraff, M. H. (1926). Corrections for chance and “guess” vs. “do not guess” instructions in multiple-response tests. *Journal of Educational Psychology*, 17, 368–375.

- Salganik, M. J., & Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34(1), 193–239.
- Sapsford, R. (2007). *Survey research* (2nd ed.). London: Sage.
- Saris, W. E. (1991). *Computer-assisted interviewing*. Newbury Park, CA: Sage.
- Saris, W. E. (1998). Ten years of interviewing without interviewer: The Telepanel. In M. P. Couper, R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nichols, & J. M. O'Reilly (Eds.), *Computer assisted survey information collection*. New York: John Wiley and Sons.
- Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, NJ: John Wiley & Sons.
- Saris, W., Revilla, M., Krosnick, J. A., & Shaeffer, E. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, 4, 61–79.
- Schaeffer, N. C. (1980). Evaluating race-of-interviewer effects in a national survey. *Sociological Methods and Research*, 8, 400–419.
- Schuman, H. (2008). *Method and meaning in polls and surveys*. Cambridge, MA: Harvard University Press.
- Schuman, H., & Converse, J. M. (1971). The effects of black and white interviewers on black responses in 1968. *Public Opinion Quarterly*, 35, 44–68.
- Schuman, H., Ludwig, J., & Krosnick, J. A. (1986). The perceived threat of nuclear war, salience, and open questions. *Public Opinion Quarterly*, 50, 519–536.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. San Diego, CA: Academic Press.
- Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys*. New York: Academic Press.
- Schuman, H., & Scott, J. (1989). Response effects over time: Two experiments. *Sociological Methods and Research*, 17, 398–408.
- Schuman, H., Steeh, C., & Bobo, L. (1985). *Racial attitudes in America: Trends*

and interpretations. Cambridge, MA: Harvard University Press.

- Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 45, 513–523.
- Schwarz, N., Hippler, H., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, 49, 388–395.
- Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, F. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, 55, 618–630.
- Sears, D. O. (1983). The persistence of early political predispositions: The role of attitude object and life stage. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 4, pp. 79–116). Beverly Hills, CA: Sage.
- Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51, 515–530.
- Shen, F., Wang, N., Guo, Z., & Guo, L. (2009). Online network size, efficacy, and opinion expression: Assessing the impacts of Internet use in China. *International Journal of Public Opinion Research*, 21, 451–476.
- Singer, E., Van Hoewyk, J., & Maher, M. P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, 64, 171–188.
- Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. *The Annals of the American Academy of Political and Social Science*, 645, 112–141.
- Smith, T. W. (1983). The hidden 25 percent: An analysis of nonresponse in the 1980 General Social Survey. *Public Opinion Quarterly*, 47, 386–404.
- Smith, T. W. (1987). That which we call welfare by any other name would smell sweeter: An analysis of the impact of question wording on response patterns. *Public Opinion Quarterly*, 51, 75–83.
- Smyth, J. D., Dillman, D. A., Christian, L. M., & McBride, M. (2009). Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion*

Quarterly, 73, 325–337.

Sniderman, P. M., & Tetlock, P. E. (1986). Symbolic racism: Problems of motive attribution in political analysis. *Journal of Social Issues*, 42, 129–150.

Sniderman, P. M., Tetlock, P. E., & Peterson, R. S. (1993). Racism and liberal democracy. *Politics and the Individual*, 3, 1–28.

Stapp, J., & Fulcher, R. (1983). The employment of APA members: 1982. *American Psychologist*, 38, 1298–1320.

Steve, K. W., Burks, A. T., Lavrakas, P. J., Brown, K. D., & Hoover, J. B. (2008). Monitoring telephone interviewer performance. In J. M. Lepkowski, C. Turner, J. M. Brick, E. D. de Leeuw, L. Japac, P. J. Lavrakas, M. W. Link, & R. L. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 201–422). New York: Wiley.

Sudman, S. (1976). *Applied sampling*. New York: Academic Press.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions*. San Francisco: Jossey-Bass.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Taylor, J. R., & Kinnear, T. C. (1971). Numerical comparison of alternative methods for collecting proximity judgements. *American Marketing Association Proceeding of the Fall Conference*, 547–550.

Thornberry, Jr., O. T., & Massey, J. T. (1988). Trends in United States telephone coverage across time and subgroups. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 25–50). New York: Wiley.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, 103, 299–314.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Traugott, M. W., Groves, R. M., & Lepkowski, J. M. (1987). Using dual frame designs to reduce nonresponse in telephone surveys. *Public Opinion*

Quarterly, 51, 522–539.

- Trzesniewski, K. M., Donnellan, B., & Lucas, R. E. (Eds.). (2010). *Secondary data analysis: An introduction for psychologists*. Washington, DC: American Psychological Association.
- Trzesniewski, K. M., Donnellan, M. B., & Robins, R. W. (2003). Stability of self-esteem across the life span. *Journal of Personality and Social Psychology*, 84, 205–220.
- Tziner, A. (1987). Congruency issues retested using Rineman's achievement climate notion. *Journal of Social Behavior and Personality*, 2, 63–78.
- United Nations Economic and Social Commission for Asia and the Pacific. (1999a). Computer Assisted Personal Interviewing (CAPI). In *Guidelines on the application of new technology to population data collection and capture*. New York: The United Nations. Retrieved August 25, 2013, from http://www.unescap.org/stat/pop-it/pop-guide/capture_ch03.pdf.
- United Nations Economic and Social Commission for Asia and the Pacific. (1999b). Computer Assisted Telephone Interviewing (CATI). In *Guidelines on the application of new technology to population data collection and capture*. New York: The United Nations. Retrieved August 26, 2013, from http://www.unescap.org/stat/pop-it/pop-guide/capture_ch04.pdf.
- Van De Walle, S., & Van Ryzin, G. G. (2011). The order of questions in a survey on citizen satisfaction with public services: Lessons from a split-ballot experiment. *Public Administration*, 89, 1436–1450.
- Visser, P. S., Krosnick, J. A., Marquette, J., & Curtin, M. (1996). Mail surveys for election forecasting? An evaluation of the Columbus Dispatch poll. *Public Opinion Quarterly*, 60, 181–227.
- Voogt, R. J. J., & Van Kempen, H. (2002). Nonresponse bias and stimulus effects in the Dutch National Election Study. *Quality and Quantity*, 36, 325–345.
- Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology*, 52, 133–142.
- Watson, D. (2003). Sample attrition between waves 1 and 5 in the European Community Household Panel. *European Sociological Review*, 19, 361–378.

- Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, 67, 1049–1062.
- Wegener, D. T., Downing, J., Krosnick, J. A., & Petty, R. E. (1995). Measures and manipulations of strength-related properties of attitudes: Current practice and future directions. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 455–487). Hillsdale, NJ: Erlbaum.
- Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27, 236–247.
- Weisberg, H. F., Haynes, A. A., & Krosnick, J. A. (1995). Social group polarization in 1992. In H. F. Weisberg (Ed.), *Democracy's feast: Elections in America* (pp. 241–249). Chatham, NJ: Chatham House.
- Weisberg, H. F., Krosnick, J. A., & Bowen, B. D. (1996). *An introduction to survey research, polling, and data analysis* (3rd ed.). Newbury Park, CA: Sage.
- Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, 64, 956–972.
- Wesman, A. G. (1946). The usefulness of correctly spelled words in a spelling test. *Journal of Educational Psychology*, 37, 242–246.
- Wikman, A., & Warneryd, B. (1990). Measurement errors in survey questions: Explaining response variability. *Social Indicators Research*, 22, 199–212.
- Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage Publications.
- Wilson, D. C. (2010). Perceptions about the amount of interracial prejudice depend on racial group membership and question order. *Public Opinion Quarterly*, 74, 344–356.
- Winkler, J. D., Kanouse, D. E., & Ware, Jr., J. E. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology*, 67, 555–561.
- Wiseman, F. (1972). Methodological bias in public opinion surveys. *Public*

Opinion Quarterly, 36, 105–108.

- Wojcieszak, M. E. (2012). On strong attitudes and group deliberation: Relationships, structure, changes, and effects. *Political Psychology*, 33, 225–242.
- Wright, S. D., Middleton, R. T., & Yon, R. (2012). The effect of racial group consciousness on the political participation of African-Americans and Black ethnics in Miami-Dade County, Florida. *Political Research Quarterly*, 65, 629–641.
- Yalch, R. F. (1976). Preelection interview effects on voter turnouts. *Public Opinion Quarterly*, 40, 331–336.
- Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and Internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75, 709–747.
- Zabel, J. E. (1998). An analysis of attrition in the panel study of income dynamics and the survey of income and program participation with an application to a model of labor market behavior. *The Journal of Human Resources*, 33, 479–506.
- Zagorsky, J., & Rhoton, P. (1999). *Attrition and the national longitudinal survey's women cohorts*. Manuscript, Center for Human Resource Research, Ohio State University.
- Ziliak, J. P., & Kniesner, T. J. (1998). The importance of sample attrition in life cycle labor supply estimation. *Journal of Human Resources*, 33, 507–530.

¹ Although we describe surveys as focused on sampling from a population of people, some surveys randomly sample from a list of organizations and seek informants at the organizations to interview.

² In principle, it is preferable for interviewers to be blind to the experimental condition to which each respondent is assigned, but this may be impractical in many instances. Therefore, it may be best simply to be sure that interviewers are unaware of the hypotheses being tested in an experiment.

³ In some surveys, the goal is to describe a population of objects (e.g., airplanes) and to use people as informants to describe those objects.

⁴ Some have argued that the requirement of independence among sample elements eliminates systematic sampling as a probability sampling method, because once the sampling interval has been established and a random start value has been chosen, the selection of elements is no longer independent. Nevertheless, sampling statisticians and survey researchers have traditionally regarded systematic sampling as a probability sampling method, as long as the sampling frame has been arranged in a random order and the start value has been chosen through a random selection mechanism (e.g., Henry, 1990; Kalton, 1983; Kish, 1965). We have, therefore, included systematic sampling as a probability sampling method, notwithstanding the potential problem of nonindependence of element selection.

⁵ Most researchers use the term “sample” to refer both to (a) the set of elements that are sampled from the sampling frame from which data ideally will be gathered and (b) the final set of elements on which data actually are gathered. Because almost no survey has a perfect response rate, a discrepancy almost always exists between the number of elements that are sampled and the number of elements from which data are gathered. Lavrakas (1993) suggested that the term “sampling pool” be used to refer to the elements that are drawn from the sampling frame for use in sampling and that the term “sample” be preserved for that subset of the sampling pool from which data are gathered.

Chapter seventeen Conducting Research on the Internet

Michael R. Maniaci and Ronald D. Rogge

Over the past several decades, the Internet has rapidly become increasingly accessible worldwide. Global Internet access has more than quadrupled in the past decade alone, covering more than a third of the global population (International Telecommunication Union, [2012](#)). In 2011, nearly 80% of adults in the United States used the Internet (Pew Internet and American Life Project, [2011](#)). Taking advantage of these trends, social scientists have begun capitalizing on the Internet as a powerful tool to collect data from large and relatively diverse samples with minimal associated costs. As Internet data collection becomes more widely adopted by social-personality psychologists, and as advances in technology allow increasingly innovative usages, examining the nature of this emerging method becomes increasingly important.

This chapter explores the promise and pitfalls of using the Internet as a tool to collect data, with a focus on practical and conceptual concerns relevant to social-personality psychologists. Although there is a wealth of research exploring Internet use as a topic of study in its own right (e.g., studies examining computer-mediated communication; Kruger, Epley, Parker, & Ng, [2005](#)), this chapter focuses primarily on using the Internet to recruit participants and collect data. We first discuss some of the benefits and challenges of collecting data over the Internet. In the second section, we discuss effective strategies for implementing a set of common study designs online. Next, we address practical issues relevant to implementing Internet-based studies, and we end with a discussion of future opportunities for Internet use in research.

Conceptual Issues

Why Collect Data over the Internet?

Collecting Large Samples. The primary benefit of collecting data on the Internet is that it allows researchers to recruit large samples of participants within a

relatively short time frame. For instance, Erdle, Gosling, and Potter (2009) examined the association between the Big Five personality factors and self-esteem among 628,640 participants recruited from the Internet. Access to such large samples has the potential to address an enduring Achilles' heel of psychological research: the low levels of power associated with small sample sizes. Studies of typical power levels in psychology journals across several decades have consistently found average power of only around .20 for identifying small effect sizes and .50–.60 for identifying medium effect sizes (e.g., Clark-Carter, 1997; Cohen, 1962; Rossi, 1990; Sedlmeier & Gigerenzer, 1989), suggesting that meaningful effects are going undetected. Small sample sizes also increase the risk of reporting false-positive findings (i.e., Type I errors), even in papers with multiple conceptual replications (Simmons, Nelson, & Simonsohn, 2011). To make matters worse, random assignment in small samples is likely to result in groups that differ markedly on extraneous variables (Hsu, 1989), potentially confounding experimental manipulations. The most effective method of addressing these problems would be to routinely collect samples with adequate power, and the Internet can provide researchers a practical method of doing just this.

A More Diverse Alternative to Student Samples. Given their low cost and convenience, undergraduate participant pools have been a mainstay of social-personality studies for decades. We examined the April 2012 issue of the *Journal of Personality and Social Psychology* (JPSP) and found that approximately 80% of the samples consisted entirely of undergraduate students. This is consistent with earlier reports that student samples were used in 85% of JPSP articles in 2002 (Gosling, Vazire, Srivastava, & John, 2004) and more than 70% of social psychology studies since the 1960s (Sears, 1986). Although this homogeneous population may be adequate for examining many basic social processes, certain phenomena (e.g., parenthood, work-place dynamics, divorce, illness, unemployment) are not well represented within student samples, and potential moderator variables (e.g., age, race, socioeconomic status, education, marital status, intelligence) have a markedly restricted range in such samples. Psychology participant pools in particular often include many more females than males.

The unrepresentative nature of student samples raises the concern that many social-personality processes that have been repeatedly observed in student samples may operate differently in other groups. For instance, mortality salience studies have significantly smaller effect sizes in non-student samples than in student samples (Burke, Martens, & Faucher, 2010). Similarly, a meta-analysis

of more than 4,500 studies examining various behavioral and psychological constructs found that nearly half of the observed effect sizes differed substantially between student and non-student samples (Peterson, 2001). Thus, the field's heavy reliance on such convenience samples could limit both the scope of our research and the generalizability of our findings (Sears, 1986).

In contrast, Internet-based research provides access to samples that are considerably more diverse.¹ Gosling *et al.* (2004) compared the demographics of more than 350,000 participants recruited online to those of participants recruited by more traditional means from all of the studies reported in *JPSP* in 2002. The Internet sample was more diverse in terms of gender, geographic region, age, and socioeconomic status. In addition to greater demographic diversity, Internet-based samples tend to exhibit greater variability than student samples on individual differences (Birnbaum, 2004).

Internet-based studies also provide a highly cost-effective method of collecting cross-cultural data. For instance, Park, Peterson, and Seligman (2006) examined self-reported character strengths in a sample of more than 100,000 adults from 54 nations and 50 U.S. states. Gosling, Sandy, John and Potter (2010) report that at least 100 participants had provided data at their research website from each of 111 countries. Thus, Internet-based samples can enhance external validity by providing greater diversity than student samples, allowing researchers to examine a broader set of processes across a wider range of cultural contexts and developmental stages.

Low Costs. The Internet offers a very inexpensive method of collecting data. There are costs associated with Internet-based research (e.g., website-hosting expenses, recruitment costs, time spent preparing webpages and recruiting participants). However, when considered in terms of the time and money invested per participant, the Internet offers one of the least expensive methods of conducting research. For example, a study developing and validating a new measure of relationship satisfaction (Funk & Rogge, 2007) recruited 6,389 online respondents at a marginal cost of 8.6 cents per participant (primarily involving advertising costs). Conducting a comparable study using traditional lab-based methods would have incurred far greater costs. Thus, Internet-based studies often require fewer resources than do even undergraduate participant pool samples.

Automating Research. The Internet also enables relatively simple automation of potentially time-consuming research tasks such as recruitment, study invitation, longitudinal follow-up, experimental manipulations, and display of

detailed and interactive feedback to participants. Although each of these tasks has long been successfully implemented in lab-based studies, many of them are fairly labor intensive, prone to human error, and carried out in ways that might introduce bias. For example, daily diary studies can require considerable person-hours to send daily surveys, track compliance, and send reminders – tasks that can be fully automated using Internet-based study platforms. Furthermore, by removing the researcher from direct contact with participants in Internet-based experiments, it is possible to prevent experimenter bias and demand characteristics that might sway participants' behavior. Internet-based methods also eliminate another source of error endemic to paper-based research: data entry errors. For these reasons, some researchers use Internet-based methods to collect data even in laboratory research. Thus, although Internet-based methods present unique challenges (discussed later in the chapter), they also hold potential to not only save time and effort, but also to increase the quality of data by minimizing certain forms of bias and error.

Reaching Rare Populations. The Internet has also facilitated the study of groups that are difficult to recruit using other methods, either because they are rare (e.g., certain clinical populations, individuals experiencing a severe stressor or major life transition) or hesitant to participate in research (e.g., stigmatized groups). For example, Meier, Fitzgerald, Pardo, and Babcock (2011) used the Internet to recruit one of the largest samples of female-to-male transsexuals to date, with more than 400 respondents, offering unique insights into this exceedingly rare population. Bogart and Matsumoto (2010) examined emotion recognition accuracy in 37 adults with Möbius syndrome, an extremely rare congenital disorder characterized by facial paralysis. This was the largest sample of such individuals collected for a psychological study at the time, representing approximately 5% of all known cases in the United States. The anonymity of the Internet may also facilitate the study of groups prone to self-presentation concerns or hesitant to participate in traditional studies (e.g., individuals who engage in deviant or illegal behavior). For instance, Glaser, Dixit and Green (2002) surreptitiously conducted semistructured interviews with white supremacists by visiting racist Internet chat rooms, examining participants' likelihood to advocate racial violence in response to different types of threats .

Support for a Broad Array of Study Designs. As technology and bandwidth have steadily improved, the Internet has become capable of supporting a broad array of methods central to social-personality psychology. The Internet is not merely a tool for collecting one-shot surveys: It may be used to conduct true experiments (e.g., Sullivan, Landau, Branscombe & Rothschild, 2012), recruit

and follow-up longitudinal samples, (e.g., Saavedra, Chapman, & Rogge, 2010), conduct experience sampling designs (e.g., Reis et al., 2010; see Reis, Gable, & Maniaci, Chapter 15 in this volume), present stimuli outside of conscious awareness, assess implicit attitudes (e.g., Nosek, Banaji, & Greenwald, 2012; Lee, Rogge, & Reis, 2010; see Gawronski & De Houwer, Chapter 12 in this volume), and even to observe the behavior of individuals and groups interacting in immersive, three-dimensional virtual environments (e.g., Frost, Chance, Norton, & Ariely, 2008). In fact, Internet methods have been used to study most topic areas in social-personality psychology, including attitudes (e.g., Greenwald, McGhee, & Schwartz, 1998), prosocial behavior (e.g., Wright & Li, 2011), social cognition (e.g., Koo, Algor, Wilson, & Gilbert, 2008), persuasion (e.g., Guadagno & Cialdini, 2002), social influence (e.g., Bagozzi, Dholakia, & Mookerjee, 2006), group processes (e.g., Amichai-Hamburger, 2005), prejudice (e.g., Binning & Sherman, 2011), social support (e.g., Trepte, Reinecke, & Juechems, 2012), romantic relationship processes (e.g., Rosen, Cheever, Cummings, & Felt, 2008), and individual differences (e.g., Marcus, Machilek, & Schütz, 2006; Soto, John, Gosling, & Potter, 2008), among many others.

As technology advances, researchers will have ever-increasing opportunities to employ novel study designs. Emerging hardware and software technology are beginning to enable the collection of rich data using sensors available on modern mobile devices. For instance, GPS, accelerometers, and gyroscopes can precisely monitor movement and activity throughout the day (e.g., Das, 2012), Bluetooth and other wireless protocols may be used to assess social proximity and patterns of interaction, and custom sensors can be used to remotely collect ambulatory physiological data such as blood pressure and heart rate (Miller, 2012) .

Challenges in Internet Research

Despite the potential of the Internet to efficiently collect large, relatively diverse samples with a variety of designs, there are also unique challenges to collecting data on the Internet.

Representativeness of Internet-Based Samples. Although Internet-based samples seem to be more diverse than college student samples (Gosling et al., 2004), they typically do not represent larger populations as well as do probability or random sampling recruitment strategies (e.g., random-digit dialing).² In part, this is simply owing to the fact that Internet samples (like student samples) are usually samples of convenience (e.g., made up of people responding to advertisements) and therefore subject to the same concerns that

apply to the vast majority of published research. Of course, Internet-based studies require that participants have access to the Internet and therefore may be biased in demographic characteristics that are correlated with Internet access, such as race, gender, geographic location, age, and income (e.g., Krantz & Dalal, 2000). Compared to the general population, Internet users tend to be younger, higher in socioeconomic status and education, and less likely to be unemployed, living in a rural area, or a member of a racial or ethnic minority group (Lenhart et al., 2003). These socioeconomic differences are particularly evident with regards to high-speed or broadband Internet access (Pew Internet and American Life Project, 2011). Individuals with disabilities are also significantly less likely to use the Internet (Lenhart et al., 2003) and may not be able to participate in studies with low accessibility (e.g., those requiring the use of a monitor or keyboard).

Although demographic differences in Internet use persist, they have gradually diminished in size as Internet access has expanded over the last decade (e.g., National Telecommunications and Information Administration, 2011), and the diversity of samples recruited over the Internet is likely to increase with expanded access (Murray & Fisher, 2002). Current estimates suggest a high degree of penetration, with 78% of the population of North America accessing the Internet (Pew Internet and American Life Project, 2011). Furthermore, even though Internet-based samples are typically not representative of the larger population, they are likely less biased in many ways than the samples typically used by social-personality psychologists (particularly undergraduate participant pools). For this reason, samples recruited over the Internet are likely to be more representative than typical college students samples (Murray & Fisher, 2002), which increases external validity and generalizability (Gosling et al., 2004; Krantz & Dalal, 2000).

Attention and Compliance. Given that the Internet affords researchers minimal control over the settings in which participants complete studies (including various forms of distraction), and do not afford the same degree of researcher oversight and monitoring as do laboratory studies, researchers have raised concerns regarding the degree to which participants devote sufficient attention and effort to completing Internet-based studies (e.g., Johnson, 2005; Meade & Craig, 2012). The global reach of Internet-based studies may also attract more nonnative speakers who have difficulty understanding the meaning of items and responding appropriately (Johnson, 2005). Finally, the decreased opportunity for experimenter oversight inherent in Internet-based studies can make it more difficult to ensure that participants follow instructions (Meade & Craig, 2012).

Internet-based study participants may be more likely than those in a laboratory to divert their attention among various activities (watching TV, listening to music, having a conversation), which could reduce the effectiveness of experimental manipulations. Such noncompliance is difficult to spot without direct oversight.

Despite the potential for decreased compliance with Internet-based research, several studies have suggested that Internet methods yield data of similar quality to that obtained using more traditional methods. Gosling *et al.* (2004) compared the responses of participants recruited through the Internet and through more traditional means, finding nearly identical levels of internal consistency within self-report scales as well as identical patterns of discriminant intercorrelations between the Big Five personality traits across the two samples. Other studies comparing self-report data collected using Internet methods and more traditional methods have found no meaningful differences with regards to internal consistency, factor structure, and other indicators of measurement quality (e.g., Brock, Barry, Lawrence, Dey, & Rolffs, 2012; Buchanan & Smith, 1999; Meade, Michels, & Lautenschlager, 2007). Furthermore, a large number of experiments have found identical patterns of results when comparing Internet samples to other samples (see Krantz & Dalal, 2000; Smyth & Pearson, 2011). Although there are data to suggest that as many as 25–45% of participants routinely skip blocks of instructions (e.g., Maniaci & Rogge, 2013a), this is true of both pen-and-paper packet studies as well as online studies (Oppenheimer, Meyvis, & Davidenko, 2009). Our own work suggests that this and other forms of inattention are no higher among Internet-based respondents than those in a laboratory proctored by a research assistant, with these groups demonstrating similar levels of compliance with study procedures (e.g., watching a video clip; completing a simple word identification task), and providing comparable patterns of correlational and experimental results. Although the data from individuals completing Internet studies seem to be of comparable quality to that obtained by more traditional methods, the decreased opportunity for oversight and monitoring with Internet-based research may make it more difficult to spot inattention and noncompliance on study tasks than it would be in a laboratory setting, increasing the importance of including manipulation and attention checks (discussed later in the chapter).

Study Designs Less Amenable to the Internet. Although researchers can use the Internet to implement a broad array of study designs, some designs are not well suited to Internet data collection. Many studies in social-personality psychology require manipulating environmental factors that cannot be controlled in an Internet-based study. For instance, Ijzerman and Semin (2009) found that

experimentally manipulating the temperature of a room influenced participants' ratings of closeness to their friends. Such a manipulation would simply not be feasible using the Internet. Similarly, designs requiring face-to-face interpersonal interactions are difficult to implement on the Internet, although it is possible for participants to interact in Internet-based studies using tools like instant messaging, virtual environments, or video-messaging. For example, Frost *et al.* (2008) asked pairs of participants to go on a “virtual date” in which they interacted via avatars in a virtual environment, finding that the interaction increased liking in a subsequent meeting relative to simply reading a profile. Although such interpersonal interactions are possible in Internet-based studies, computer-mediated forms of communication differ from face-to-face interaction in several ways, such as greater strategic self-presentation (Walther, 1996) and greater ambiguity resulting from limited nonverbal cues (e.g., voice tone, facial expressions; Kruger et al., 2005).

Studies of physiological processes are also generally ill-suited for Internet data collection. Participants can self-report certain biomarkers, such as height, weight, and waist circumference (Avendano, Scherpenzeel, & Mackenbach, 2011). However, directly measuring hormones such as cortisol or oxytocin requires the collection of saliva, blood, or other biological samples. A few studies have recruited participants from the Internet and then obtained biological samples through the mail, examining nicotine metabolites and DNA (Etter, Neidhart, Bertrand, Malafosse, & Bertrand, 2005), as well as measuring levels of cholesterol and salivary cortisol (Avendano et al., 2011). However, collecting such samples in a laboratory setting allows researchers to control for additional contextual factors (e.g., time of day, eating) that can affect physiological measurements, or to precisely time measurements to coincide with a specific task. Furthermore, many physiological processes must be measured using specialized equipment (e.g., fMRI) that cannot be practically implemented via the Internet.

Studies of live behavioral observation (see Heyman, Lorber, Eddy, & West, Chapter 14 in this volume) may also be more suitable for traditional laboratory settings. Although it is possible to collect video data using a participant's webcam, this engenders technical problems (e.g., low resolution, poor or inconsistent lighting, narrow field of view), particularly when attempting to capture the behavior of more than one individual. These limitations on behavioral observation coincide with recent criticism that social-personality psychology has increasingly focused on self-reports, hypothetical vignettes, and reaction time tasks rather than the direct observation of behavior (Baumeister,

Vohs, & Funder, 2007). Thus, although the Internet offers distinct advantages to researchers, it includes limitations best addressed by conducting Internet-based studies in conjunction with laboratory-based research.

How to Create Studies on the Internet

The following section details practical guidance for implementing online studies.

Hosting the Study Webpages

The first question to address for an Internet study is where it will be hosted – specifically, which Web server (computers that deliver content over the Internet) will hold the actual webpage files of the study. If researchers have direct access to servers maintained by their academic institution or company, then using those servers may provide the greatest control with little or no direct costs. Researchers can also set up their own computers as Web servers, although this requires greater technical expertise and continued effort in maintaining the servers. It is also possible to purchase storage space on servers from professional Web-hosting services, providing an inexpensive alternative without the ongoing server maintenance requirements. Finally, studies can be hosted using online study-hosting services, which have been developed specifically for creating Internet-based studies (e.g., surveygizmo.com, qualtrics.com, keysurvey.com, hostedsurvey.com, surveymonkey.com). For additional information about setting up a Web server or hosting studies with external companies, see Fraley (2004).

Creating the Study Webpages

The second question to address is how to create the study webpages. In the following section we describe the three main options for this process: programming them yourself, programming them with the help of software, or using an online study-hosting service.

OPTION 1: Programming Webpages Yourself – HTML. The least expensive option to create study webpages is to program those pages from scratch. However, we advise avoiding this option as it requires considerable technical skills and could involve writing thousands of lines of code. Most websites are created using HyperText Markup Language (HTML),³ a document format that instructs Web browsers how to present the text, links, and media that comprise a webpage. HTML operates by embedding commands within text and denoting

those commands (or “tags”) by placing them within angle brackets. For example, surrounding a phrase with HTML tags as follows would result in the text contained within being bolded on the webpage: `bolded text`. Regardless of where a study is developed or hosted, researchers can benefit from learning a few basic HTML tags that allow control of formatting (e.g., bold, italics, centering). By itself, HTML provides limited functionality (e.g., it cannot be used to randomize page order or calculate scores) and requires other types of programming to store participants’ responses. Thus, most Internet-based studies demand the use of additional programming languages in addition to HTML to provide these additional functions.

Server-Side Programming. Because of the limitations of HTML, many websites supplement HTML with additional *server-side programming*, which allows a Web server to interact more directly with a Web browser (e.g., saving responses across multiple pages, dynamically generating webpages). CGI and PHP are two common server-side programming languages used to take data from HTML forms and save the data to databases on the server hosting the study. CGI and PHP can also be used to customize webpages based on a participant's previous responses or an experimental condition. Fraley (2004) provides detailed advice for implementing Internet-based studies using server-side programming.

Client-Side Programming. JavaScript, Java, and Adobe Flash are examples of *client-side programming*, in which code is downloaded and run directly from a participants’ computer (rather than running it on the server). In contrast to server-side programming, these client-side languages tend to be easier to implement, as the code can simply be added to the study web pages. As with server-side programming, these languages can be used to add functionality and dynamic content to Internet-based studies that is not possible with HTML alone. For example, JavaScript may be used to randomly assign participants to different experimental conditions, randomize the order of questions, effect “branching” (redirecting participants to different parts of a study based on their responses), “pipe” answers from one question to another (e.g., inserting a family member's name into subsequent questions), present interactive questions (e.g., sliding response bars), record the respondent's activity on a page (e.g., cursor movement, time spent on a particular page), check for missing answers, and provide individualized feedback based on participants’ responses. Furthermore, because it is run from the participants’ computer, client-side programming can provide more precise timing (e.g., for presenting stimuli subliminally or for assessing reaction times) than server-side programming.

JavaScript is arguably one of the most accessible and widely used forms of client-side programming as it is fairly straightforward to learn, can be added directly to webpages, and does not require expensive authoring and compiling software (as is the case for other programming languages, such as Flash and Java). JavaScript should not be confused with Java, which is a more advanced programming language that allows a website to run a stand-alone program (known as a “Java applet”) within the page. These Java applets may be used for various purposes, such as presenting stimuli subliminally or measuring reaction times (as with the IAT). Other programming languages, such as Adobe Authorware or Adobe Flash, may be used for similar purposes. Despite their potential for enabling more advanced functionality, programming languages like Java and Flash typically require more advanced programming skills, require the purchase of costly software, and incur the risk that specific Internet users might not have the necessary software “plug-ins” installed on their Web browser to run that programming.

Along these same lines, the use of client-side programming can incur basic compatibility issues as some operating platforms do not support specific client-side programming languages (e.g., Apple iPads currently cannot display Flash programming) – a problem that may be exacerbated by the increasing tendency to access the Internet using smartphones, tablets, and other mobile devices. All forms of client-side programming can be disabled by Internet users and some users may drop out of a study upon seeing a Java applet loading. Stieger, Göritz, and Voracek (2011) found that 22–26% of respondents dropped out on a page including a Java applet, approximately three-to-four times more than on the preceding pages. Furthermore, dropouts related to the Java applet differed across demographic groups, with higher dropout rates among women and older participants. Thus, it might be safer to restrict oneself to the use of HTML and more universally compatible programming like JavaScript (although a small percentage of Internet users disable JavaScript as well). Despite these limitations, client-side programming remains a broadly used method for making websites more interactive and dynamic, and as a result the vast majority of Internet users have the necessary plug-ins installed to accept such programming.

OPTION 2: Software that Facilitates Internet-Based Study Design. Although understanding the technical underpinnings of website design can enable a researcher to create more advanced and flexible Internet-based studies, these skills are by no means necessary for collecting data on the Internet. There are several free and commercial programs that simplify the creation of Web-based studies, while still offering the flexibility afforded by hosting the study on one's

own web server. For instance, many commercial programs generate HTML code based on a graphical user interface (e.g., Dreamweaver). Other software suites are more specifically focused on the needs of researchers. For instance, LimeSurvey (www.limesurvey.org) is a free and open-source software application that allows users to create studies to be hosted on a Web server. Other programs facilitate the development of Internet-based experiments with random assignment. For instance, WEXTOR is a free tool that uses a graphical user interface to generate HTML and JavaScript code needed for experimental designs with random assignment, including factorial designs (Reips & Neuhaus, 2002). Inquisit Web (www.millisecond.com) and PxLab (www.pxlab.de) both allow the creation of complex experimental designs including reaction time measurement, although both also rely on Java and may incur the compatibility issues described previously. These software suites can notably simplify the programming of a study website, but they still involve a learning curve and a moderate level of technical skill to implement.

OPTION 3: Internet-Based Study Hosting Services. By far the most user-friendly method of implementing a study on the Internet is to use a study-hosting service, which would not require programming skills or other technical knowledge. Such services are typically designed to be exceptionally user-friendly and require virtually no technical understanding of HTML, Web programming, or server technology. In addition, many of these services offer greater security and stability (e.g., minimizing unavailability attributable to technical problems) than researchers could offer by programming their own study website. There are literally hundreds of such services where researchers can create studies that are hosted by the company itself (for a comprehensive list see www.websm.org). Although these services change rapidly, a few popular services at the time of writing include SurveyGizmo (www.surveygizmo.com), SurveyMonkey (www.surveymonkey.com), LimeService (www.limeservice.com), ProtoGenie (www.protoenie.com), and Qualtrics (www.qualtrics.com). These services differ widely in their pricing, capabilities, and user interfaces. Given the large number of options available and the fact that Internet study services change rapidly, we will not catalog current features or pricing of specific services. Rather, we will highlight a few critical issues that researchers should consider when selecting a service.

Questions You Should Ask When Selecting an Internet-Hosting Service

How Much Will the Service Cost? Pricing varies widely (ranging from free to hundreds of dollars per month), and can end up being a primary factor in a researcher's decision. Hosting services may charge per response or a monthly or annual service fee for unlimited use. Some services offer site licenses, which allow a school or department to purchase access for multiple individuals for a discounted fee. Even with individual licenses, some services offer substantial discounts for academic or student use. Many services offer different pricing tiers, which vary in functionality and limits on numbers of questions, studies, and responses.

Can I Customize Formatting? Some services allow only minimal customization of fonts, colors, and formatting, whereas others provide greater flexibility in formatting. The greatest flexibility is offered by services that allow customization of formatting using HTML or Cascading Style Sheets (CSS). As the quality of formatting can affect response rates (Edwards et al., 2002), the ability to produce a highly professional-looking Internet-based study may improve data quality. In addition, not all hosting services can accommodate a broad range of question types, such as semantic differential scales (items that are made up of opposing word pairs), the use of images as response options, or other more novel item types. Some study hosting services also allow researchers to customize the flow of a survey (e.g., branching participants to different pages based on their responses or a randomly assigned condition) .

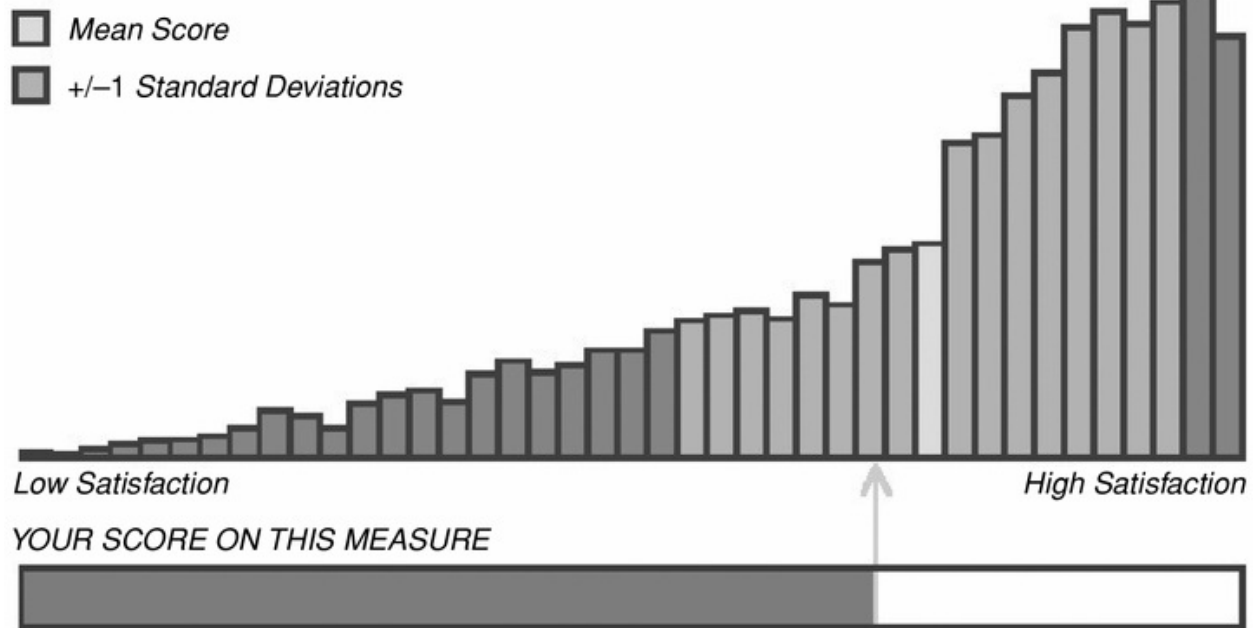
Can I Add Client-Side Programming? Some study hosting services (e.g., Qualtrics, SurveyGizmo) allow researchers to add JavaScript code or other custom programming to their studies whereas other services (e.g., SurveyMonkey, at the time of writing) prevent such additions. This can be an important factor as JavaScript or other client-side programming is often necessary to add study-specific functionality (e.g., creating sliding response scales, tracking mouse movements, setting time limits, randomly assigning participants to different experimental conditions, calculating scale totals, and adding other interactive elements). Thus, actively disabling such additions can severely restrict researchers' ability to tailor an Internet-based study to their specific needs.

Can I Program the Study to Give Immediate Feedback? A more novel feature offered by some study-hosting services is the ability to program a study to calculate totals and then provide respondents with immediate, personalized feedback – for example, comparing an individual's scores with normative data from other samples. SurveyGizmo is one of the few hosting services to offer this

functionality at the time of this writing. However, individualized feedback can be added to other services using JavaScript programming, which can also allow text feedback to be augmented with graphics (e.g., presenting a histogram showing where an individual's score lies in the overall distribution of scores on each scale; see [Figure 17.1](#)). In the authors' experience, providing such feedback can serve as a powerful recruitment incentive. Similar programming can also be used as an experimental manipulation (e.g., presenting false feedback or showing responses for a hypothetical person that are either similar or dissimilar to one's own).

RELATIONSHIP SATISFACTION

DISTRIBUTION OF SCORES ON THIS MEASURE



YOU ARE REASONABLY HAPPY IN YOUR RELATIONSHIP - Your answers place you right at the average level of relationship satisfaction when compared to the thousands of individuals who have completed this measure.

Figure 17.1. Sample individual feedback.

Can I Prevent Participants from Changing Answers? Many services allow researchers to prevent participants from moving backward through a study so that they cannot change answers once a page is submitted. This is a particularly important feature for eligibility questions, as it can be used to prevent ineligible participants from going back to change answers on eligibility questions. It can also be used to prevent individuals from changing answers on prior scales after being exposed to an experimental manipulation or gaining insight into the

study's purpose.

Can Participants Complete a Study in More than One Sitting? When implementing a longer (e.g., > 30min) study online, it might not be realistic to expect all participants to complete it in one sitting. Some study hosting services (e.g., SurveyGizmo, SurveyMonkey) can be configured to allow individuals to exit a study and return to finish it later on the same computer (picking up where they left off). A smaller number of hosting services (e.g., HostedSurvey.com, KeySurvey.com) allow participants to create login IDs with passwords so that they can complete a study in as many sittings as necessary from any computer or device. However, study-hosting services providing that level of functionality also tend to be more expensive .

How Will I Access the Data? Services differ in how they provide access to data. Most services allow data to be downloaded in spreadsheet format, but these files may require extensive cleaning (e.g., typing in names for each variable) before being imported into statistical programs. Other services allow data to be downloaded directly in the format of popular statistical programs (like SPSS), which can save considerable time and effort – particularly if the data set comes well annotated. Both SurveyGizmo and Qualtrics, for example, add the exact item wording as a variable label and the response options as value labels for each variable in an SPSS data set, which may save time and reduce errors in cleaning and scoring responses.

Will Data Security be Adequate for My Needs? Security procedures can be important for maintaining the confidentiality of participants' responses. Any information transmitted across the Internet (including email and online financial transactions) carries some risk of being intercepted by an unintended third party. To minimize this risk, some websites (including many study-hosting services) encrypt the information before transmitting it between the user's computer and the Web server (typically using HTTPS) to make it more difficult for third parties to read the information. Some study-hosting services offer HTTPS encryption as a standard or optional service, but encryption is not always available and can be difficult to implement when hosting studies on one's own Web server. Study-hosting services also differ in their physical security procedures (e.g., storing data on encrypted hard drives; securing servers in locked rooms with limited access). In some cases, researchers may need to determine if a study-hosting service meets specific institutional or legal requirements regarding security and privacy practices. For instance, researchers working with health care providers in the United States to collect personal health

information may require a HIPAA-compliant service. Although many study-hosting services (e.g., SurveyGizmo, Qualtrics) follow HIPAA guidelines, other services (e.g., SurveyMonkey, at the time of writing) do not follow HIPAA guidelines and therefore specifically prohibit the collection of personal health information by HIPAA-covered entities.

Of course, concerns about data security and storage procedures are not equally important for all types of research. Although anonymous studies or studies addressing relatively benign topics may not benefit from the additional complexity and cost of data encryption, researchers should examine the availability and cost of security features that are appropriate for their needs.⁴

Specific Internet Study Designs

In this section, we provide more detailed examples of how common study designs within social-personality psychology can be successfully implemented via the Internet.

Implementing Experiments Online

Although the earliest Internet-based studies used simple survey designs, more than half of the studies included in Skitka and Sargis's (2006) review of Internet-based research used experimental methods. Using an experimental design in an Internet-based study requires translating the random assignment process and the experimental manipulation(s) to a Web-based format.

Random Assignment Online. Although HTML does not provide the ability to randomly assign participants to different conditions, this can be accomplished with the addition of either client-side or server-side programming. For example, a few lines of JavaScript code could generate a random value of 0 or 1, which could then be used to determine which version of a manipulation or counterbalanced order a participant experiences (e.g., routing them to different web pages based on that value). Some Internet-based study hosting services have built-in random assignment functions that do not require additional JavaScript programming.

Text Manipulations. The simplest implementation of an experiment on the Internet involves exposing respondents to different sets of text instructions or stimuli to manipulate a construct of interest. For example, Oppenheimer *et al.* (2009) demonstrated a sunk cost effect in that participants reported being more

likely to attend a sports event on a cold day if the scenario described them having paid for a ticket as opposed to receiving a free ticket. Researchers should be cautious about using lengthy text (e.g., a small change in wording within a long paragraph) as the primary manipulation for a variable. In the same study, Oppenheimer and colleagues found that approximately 35–45% of participants failed to read instructions or descriptive text carefully, and that the effects of their text manipulation were stronger after excluding those inattentive participants. As a result, we suggest that researchers make their manipulations as engaging as possible by augmenting text with pictures, placing the experimental manipulation on a page all by itself to encourage more careful reading, and using manipulation checks to ensure compliance.

Dynamic/Interactive Manipulations. Given the increasing multimedia capabilities of the Internet, researchers can also use interactive tasks to manipulate independent variables. For example, Williams, Cheung, and Choi (2000) experimentally manipulated ostracism by having participants play a virtual ball-tossing game (“Cyberball”) with two other animated avatars. The avatars (controlled by the computer) gradually stopped tossing the ball to participants in the ostracism condition. The researchers found that this simple manipulation influenced mood, feelings of control, sense of belonging, and likelihood to conform on subsequent tasks.

Although the Cyberball manipulation required extensive programming, programming knowledge is certainly not necessary to manipulate independent variables via the Internet. For example, G6ritz (2007) successfully manipulated mood in a series of Internet-based studies by showing participants pictures and cartoons designed to elicit either positive or negative mood. If a researcher wanted to use video clips to manipulate mood (as described by Quigley, Lindquist, & Barrett, Chapter 10 in this volume), he or she could upload them to YouTube.com and then have the YouTube website generate the code necessary to display each clip in the study's webpages.

Implementing Indirect Measures and Tasks Online

Another set of paradigms commonly used by social-personality psychologists involve presenting stimuli for very precise amounts of time in order to activate a concept (e.g., priming; see Bargh & Chartrand, Chapter 13 in this volume) or to measure attitudes of which participants might not be fully aware (e.g., implicit measurement; see Gawronski & De Houwer, Chapter 12 in this volume). Both types of tasks typically require precise timing, either in the presentation of

stimuli or in the measurement of reaction times. Such tasks may require a keyboard and are typically not compatible with mobile devices – something that should be made clear to participants. Also, Internet-based response time measurement may entail slightly greater error than a comparable lab study, as response times can be influenced by variability in participants' hardware (e.g., monitor refresh rates, polling rates of keyboards and mice).⁵ There are, however, a few options for implementing priming and reaction time tasks on the Internet with reasonable precision.

Inquisit 3 Web. One commercial option is Inquisit 3 Web (www.millisecond.com), a software platform that enables online administration of reaction time tasks like the Implicit Association Task (IAT; Greenwald et al., 1998) and the Go/No-go Association Task (GNAT; Nosek & Banaji, 2001) as well as other interactive tasks (e.g., Wisconsin Card Sorting Test, Iowa Gambling Task). Using a comprehensive platform such as Inquisit 3 Web provides considerable flexibility while requiring relatively little programming knowledge. However, Inquisit 3 Web has two notable disadvantages. First, it uses Java Network Launching Protocol (JNLP), installing a small Java program on each participant's computer in order to run the actual study tasks. As a result, when participants reach the task webpage, they will typically receive a pop-up alert from their browser asking for permission to download and run the Java program, often with a warning (e.g., “This type of file can harm your computer”). Such barriers can dramatically increase study dropout (e.g., Stieger et al., 2011) and introduce compatibility issues (e.g., Inquisit 3 Web is not compatible with Mac OS, although this may change in future releases). A second drawback is that Inquisit 3 Web is expensive, with a single annual license (which may be used to run only a single study at a time) currently costing nearly \$1,500.

PXLab. PXLab (www.pxlab.de) is a free programming package that facilitates the creation of Internet-based studies with reaction-time tasks. Like Inquisit 3 Web, PXLab uses Java programming, which means that participants typically need to view and accept a warning message before the program will run on their computer. PXLab also requires hosting files on a Web server and necessitates programming to configure and embed in an Internet-based study. PXLab is, however, compatible with various computers and operating systems, including Mac OS.

Implementing Longitudinal and Experience Sampling Studies

Another common set of social-personality paradigms involves collecting follow-up data to predict change over time, or to examine within-person fluctuations in processes over time or across contexts. For example, Saavedra, Chapman, and Rogge (2010) provided support for negative conflict behavior and mindfulness as moderators of the longitudinal effects of attachment insecurity among 865 online respondents followed on a monthly basis for 1 year. The Internet may also be used with daily diary methods and other experience sampling or intensive longitudinal designs (see Reis, Gable, & Maniaci, Chapter 15 in this volume). Park, Armeli, and Tennen (2004) used the Internet to collect daily diary data on appraisals and coping, asking participants to report on daily stressful events, appraisals of controllability, coping strategies, and mood for 28 days. The following section provides practical advice for collecting longitudinal and experience-sampling data via the Internet.

Collecting Contact Information. It is typically necessary to collect some form of contact information in order to invite participants to complete follow-up assessments. With Internet-based studies this can be as simple as obtaining an email address – offering a significant advantage over traditional recruitment strategies that may require participants to share their full names and mailing addresses. We recommend asking for personal email addresses in addition to work-or school-based email accounts as the latter tend to become inactive as people take new positions. If longitudinal data is critical to a study's main hypotheses, then the study could require that participants share their email addresses at the beginning of the study; however, requiring email addresses tends to increase dropout rates (O'Neil & Penrod, 2001). In our personal experience, when offered a choice on whether to provide their email address or not, roughly 50–70% of participants will do so and then roughly 50–80% of these individuals will provide longitudinal data when invited (dependent on incentives, study length, and other factors). Attrition can be relatively high in Internet-based longitudinal studies, which at a minimum necessitates analyses to determine if there is bias in attrition (see Mazza & Enders, Chapter 24 in this volume). If participants who provide follow-up data differ from those who do not, including relevant variables as covariates in analyses can reduce the potential bias caused by attrition (Singer & Willett, 2003, p. 158).

Linking Waves of Longitudinal Data. Some of the more expensive study hosting platforms (e.g., HostedSurvey, KeySurvey) include the option of creating log-in accounts for each participant – providing a method of linking data from the same participant across time. It is also possible (albeit more difficult) to program longitudinal tracking with log-ins using server-side

programming in conjunction with a database (e.g., using PHP with a MySQL database). Another option for linking data across assessments is to ask participants to provide their email address near the beginning of each follow-up survey. This generally works quite well; however, email addresses can change over time. If a researcher is relying on email addresses alone to link data, then he or she is likely to have a small percentage of follow-up responses that cannot be directly linked.

A far more reliable method of tracking data from participants across time is to embed a unique, researcher-assigned ID number within the link to the survey provided in each invitation email. This approach simply requires adding a code to the end of the link (e.g., "...index.html?id=xxxx" where "xxxx" would be replaced with each participant's ID number). Many online study-hosting services can automatically pull such information out of the link and save it with the rest of a participant's responses. It is also possible to capture this information by adding lines of JavaScript code to the first page of a study. Although embedding ID numbers in links is a reliable method of matching longitudinal data across assessments, it is not completely foolproof. If participants choose to manually type the study link into a browser rather than simply clicking on the link, they may fail to include the embedded ID, inadvertently de-identifying their follow-up data. Thankfully, this is rare, and it is often possible to use other identifying information (e.g., demographic information or IP addresses) to match the data. Given that any of these methods can fail, we recommend using more than one method to ensure that all follow-up responses can be matched to earlier assessments .

Automating Invitations. Another option available at many study-hosting services is the ability to automate and manage email invitations for a study. Although study invitations can be sent manually through an email program, some hosting services simplify this task and provide added functionality. For instance, emails may be programmed in advance to go out at particular times, reminder emails can be automatically sent to participants who have not yet responded by a certain time, and study links can be programmed to include participant ID numbers. Some hosting services allow multiple invitations to be scheduled in advance. These features can save considerable time in daily diary and other experience sampling methods, as multiple email invitations (and reminders) can be scheduled in advance for multiple participants, rather than having to send email invitations individually. For instance, a researcher could program daily diary surveys to be sent by email at the same time each night for a period of several weeks, with reminder emails sent to noncompliant participants

early the next morning. Although the emails from a study hosting service may have a slightly higher risk of being marked as “spam” or placed in a “junk mail” folder, invitations emailed directly from the researcher carry similar risk. As a result, researchers should consider additional steps to ensure that email invitations reach participants (e.g., by sending an initial test email and asking participants to respond to it, by contacting noncompliant participants, and by including a member of the study team in the invitation pool to ensure that emails are being sent on time).

Using Multiple Contacts for Each Wave of Follow-up. Survey invitations yield higher response rates when respondents are contacted multiple times (e.g., Dillman, 2007; Edwards et al., 2002). Göritz and Crutzen (2012) meta-analyzed response rates across 38 studies with a total of more than 240,000 participants, finding that response rates increased by an average of 16 percentage points following a single reminder. Research also suggests that closely timed reminders are more effective than reminders sent using longer delays. For example, Crawford, Couper, and Lamias (2001) found that sending a reminder after two days yielded a stronger response rate than sending a reminder after five days. We routinely use at least three to five invitations per participant for each wave of a longitudinal follow-up, sending another reminder every few days.

Offering Briefer Options for Follow-up Participation. Study length is one of the strongest factors affecting participation rates (e.g., Edwards et al., 2002; Fan & Yan, 2010), with shorter studies showing notably higher response rates. We typically strive to keep follow-up assessments within 10–15 minutes each in an effort to minimize participant burden. Follow-up survey length is particularly important in studies with multiple waves of follow-up and in everyday experience studies, as participant fatigue can increase attrition over time. Despite these efforts at reducing burden, even a 10-minute follow-up might feel too onerous for some participants. As a result, we typically prepare a brief version of the follow-up survey, restricted to just a few measures of constructs that are most central to our hypotheses in a three-to five-minute survey. We sometimes also prepare an extremely brief version including fewer than 10 questions to assess 1 or 2 critical outcomes in under a minute. Then, for example, the fourth and fifth invitation emails can include a link to the full follow-up assessment as well as a link to the brief version, offering participants an alternative option with lower time demands. We follow this email with a final email invitation to that wave of assessment that contains only the very brief version in text format, asking participants to simply reply to the email with their answers.

Building Rapport. Actively maintaining a positive relationship with participants can also help to reduce attrition rates (e.g., Dillman, 2007). For example, it is possible to send thank you, birthday, holiday, or anniversary eCards to participants via email either through online services (e.g., www.bluemountain.com, www.hallmark.com) or by simply sending a personalized email. Such small gestures can go a long way to building rapport, making participants more likely to comply with follow-up surveys. It is also possible to create a study webpage providing regular updates on a study's progress, thereby encouraging participants to invest greater levels of personal ownership in their participation in the study. Although researchers must be careful not to share information that might compromise future responses, a study update that is visually engaging and interesting can help commit participants to the project (see www.couples-research.com for an example) .

Study Design Considerations

In many ways, designing an Internet-based study presents many of the same challenges and concerns that researchers face when designing a lab-based counterpart of that same study. The following section highlights a few unique options and challenges of Internet-based study design. For a more detailed discussion of many of these topics, see Gosling and Johnson's (2010) edited book on Internet methods .

Ordering of Questions

Survey researchers have developed several strategies regarding question ordering to optimize studies (e.g., Dillman, 2007), and these also apply to Internet-based research.

- *Place eligibility questions near the beginning of the study.* This allows ineligible individuals to be screened out of the study quickly, obviating wasted time. The study website can also be programmed to automatically display a message informing participants that they are ineligible and preventing them from going backwards in the survey to change their answers. Including just a few brief questions assessing demographics or other key outcomes before such a message would help the researcher later determine if participants deemed ineligible differed on other factors.
- *Place several questions directly relevant to the topic of the study on the first*

page. Putting relevant questions early increases response rates (see Edwards et al., 2002 for a review). This can reinforce participants' trust in the researcher. For example, if a study is advertised as the "Attitudes about Sex" study, but participants are asked to complete five minutes of questions assessing ancillary constructs prior to answering a single question about sex, then respondents might feel duped or misled and may drop out before completing the study. This may be especially problematic with Internet-based studies, which tend to have relatively high dropout rates (e.g., O'Neil & Penrod, 2001).

- *If possible, randomize the order of item presentation.* Randomizing the order of items is difficult to do in pen-and-paper packets but can be accomplished easily in Internet studies. In addition to randomizing item order within a page, some hosting services allow randomizing the order of entire scales or web pages. Although researchers have largely ignored the impact of order effects on self-report data, there is evidence that order effects may influence responses (e.g., Couper, Conrad, & Tourangeau, 2007). Internet-based studies offer a distinct advantage in addressing this potential problem (see Visser, Krosnick, Lavrakas, & Kim, Chapter 16 in this volume).

Number of Questions per Page

As Internet-based studies are not tied to the limitations of an 8.5-by-11 page, a researcher is free to decide how many questions to include on each page. This means that it is possible to put critical questions, instructions, or manipulations on pages by themselves. Although all questions could appear on a single page, seeing a great many questions on a single page can discourage individuals from completing a study. Generally, putting no more than 20 questions on a survey page minimizes the need for scrolling in browsers while at the same time minimizing excessive repetition of instructions and answer choices. Studies designed for presentation on mobile devices may require even fewer questions per page.

Formatting for Presentation across Diverse Platforms

With the increasing popularity of smartphones, tablet computers (e.g., iPads), and laptops with small screens (e.g., netbooks, ultrabooks), researchers should consider how their studies will look and perform on different devices. A study might look good on a desktop computer with a 20-inch screen, but when

rendered on smaller screens, item text and response options might be cut off or fail to display properly, introducing error or encouraging dropout. In addition, if the study makes use of client-side programming such as Flash or Java, browser incompatibility could render the study nonfunctional on mobile devices. Researchers might also consider the degree to which a study is accessible to persons with disabilities. For instance, a simple text-based survey may be accessible to the visually impaired using a screen reader, whereas a study using Java programming or requiring the use of a mouse would be significantly less accessible. Thus, researchers implementing studies online should design the study to maximize compatibility across platforms and accessibility.

At a minimum, a study should be visible on a browser window no greater than 800–900 pixels wide to accommodate a broad range of desktop monitors and laptop screens. We also recommend that researchers routinely repeat response options every 6–10 rows when presenting large blocks of items sharing a common response set. This ensures that response options will remain visible for participants who are completing the study in a small window. Some online study-hosting services (e.g., SurveyGizmo) detect the use of mobile devices and allow researchers to set up different formatting for such devices.

Professional Presentation

Surveys tend to have higher response rates when presented in a polished and professional manner (e.g., Dillman, 2007; Edwards et al., 2002). In our experience, a common criticism lobbied at Internet-based studies during the recruitment phase is to question the integrity and design of the study – quite literally to accuse the study of not being “real” research. We have found several strategies that seem to decrease the rate of such negative responses. For instance, a study may emphasize the academic affiliation of the researchers by hosting on a university server and including a banner image with the name of the university on the first page. This approach has an added benefit, in that academic affiliations tend to produce higher response rates than do corporate affiliations (e.g., Edwards et al., 2002). We also find that highlighting the academic position of the principal investigator and providing contact information for him or her on the first and last pages of the survey can help direct concerns into emails sent to the researcher (as opposed to public forum posts). Finally, Internet-based studies should have a professional formatting style that is used consistently throughout the study.

Making Questions Mandatory

Most study-hosting services allow researchers to make answers to specific questions mandatory, causing a warning message to appear if a participant tries to go to the next page before providing an answer. Although this is essential for certain questions (e.g., eligibility questions or branching questions where the participants' response determines subsequent questions), we strongly caution against overuse of this functionality. Setting the majority of questions on a survey as mandatory can result in a frustrating experience for participants and may encourage dropout. Requiring answers also raises ethical concerns if participants are assured that they can skip any question they do not wish to answer. Researchers may therefore wish to reserve this functionality for only the most critical questions.

Measuring Attention

Although Internet methods have been shown to yield data of similar quality to those obtained from other methods, identifying invalid or inattentive responses is still a concern. Because of the decreased control and oversight, increased opportunities for distraction, and greater likelihood of recruiting nonnative speakers inherent in Internet-based studies, some authors have argued that assessing inattention is particularly important on the Internet (e.g., Johnson, 2005; Meade & Craig, 2012). Inattentive responding (e.g., selecting an answer choice without fully reading or considering the item) can contribute to error variance, moderate correlational and experimental results, influence factor analyses, and reduce power (Maniaci & Rogge, 2013a; Meade & Craig, 2012; Woods, 2006).

Several strategies exist for identifying inattentive respondents. For instance, Oppenheimer *et al.* (2009) developed the Instructional Manipulation Check (IMC) to identify participants who fail to read instructions. The IMC embeds important instructions (asking participants to click on a page title rather than the “continue” button at the bottom of the page) at the end of a lengthy paragraph. The IMC moderated the effectiveness of text-based experimental manipulations in two separate samples, such that the manipulations replicated previous findings among attentive participants, but not among participants identified as inattentive. The IMC identifies a large proportion of participants (approximately 35–45%) as potentially failing to read instructions. This figure far exceeds other researchers' estimates that approximately 10% of respondents engage in problematic levels of inattentive responding (Meade & Craig, 2012), making the IMC impractical as a

screening tool for identifying inattentive participants.

Inattentive responding to self-report items (as opposed to skipping instructions) can be measured using various types of indirect questions, an approach typically used in large (300–400 item) clinical assessment inventories but rarely used by social-personality psychologists. Meade and Craig (2012) describe and assess several indicators of inattentive responding, including measures of inconsistent responding (e.g., consistency across even and odd numbered items within a scale), response patterns (e.g., the longest string of identical answers), and multivariate outliers. Their results indicated that 10–12% of an undergraduate sample belonged to a latent class characterized by highly careless responding. Meade and Craig (2012) conclude that “it is imperative that Internet survey data be properly screened for careless responses in order to safeguard the integrity of research conclusions” (p. 453).

Maniaci and Rogge (2013a) built on this work by developing the Attentive Responding Scale (ARS), which identifies a small proportion of participants (typically 4–8%) exhibiting high levels of inconsistent responding (i.e., giving different answers on pairs of nearly identical items presented in opposite halves of the survey) and/or atypical responding (e.g., giving an answer of “Very True” to the question “I enjoy the music of Marlene Sandersfield,” a fictional character). Across several studies, Maniaci and Rogge found that compared to attentive respondents, participants identified as excessively inattentive by the ARS were less compliant with study tasks (e.g., watching a short video, completing a simple task) and provided responses with markedly lower internal consistency. Inattentive responding moderated the results of experimental manipulations and correlational analyses, such that prior results were replicated among attentive but not among inattentive respondents. Furthermore, screening out inattentive respondents using the ARS increased statistical power by more than 5% for two different experimental manipulations. These results suggest that it is possible to effectively measure and screen out excessively inattentive respondents, thereby noticeably improving the quality of research.

Engaging Attention

The increasing multimedia capabilities of the Internet afford the opportunity to make the questionnaire portions of Internet-based studies more dynamic and interactive than can traditional pen-and-paper measures. Adding images or brief video clips throughout a study can make it more visually engaging, helping maintain participants’ attention. In the traditional survey literature, simply using

colored ink in contrast to just black-and-white printing notably increased survey participation rates (see Edwards et al., 2002). It is also possible to incorporate interactive elements (e.g., sliding-bar questions, pictorial response options) to provide a novel, more engaging method of responding to questions. For example, following Le, Moss, and Mashek (2007), we programmed an analog version of the Inclusion of Other in the Self Scale (IOS; Aron, Aron & Smollan, 1992) into our Internet-based studies. JavaScript code presents the IOS as a slider that can take on values between 0 and 100. As participants click on and move the slider, a pair of circles labeled “Self” and “Partner” move above the slider, pictorially representing the degree of self-other overlap based on the slider's position and thereby providing engaging visual feedback.

It is also possible to program a webpage to detect inattention and provide a warning to inattentive respondents to help them reengage the study. For example, the developers of the IMC recommended that instead of screening out the 35–45% of the sample failing to carefully read the IMC instructions, the IMC page might be programmed so that it is impossible to move to the next page without reading the instructions (that is, requiring participants to click on the page title before moving forward in the survey). They demonstrated that adding such a page to a survey increased attention on subsequent pages, improving effect sizes for text manipulations (Oppenheimer et al., 2009). Thus, by using a few simple tools, it is possible to make studies more dynamic and engaging for participants and to more actively address inattention.

Study Length

As with traditional studies, it is more difficult to successfully recruit and obtain complete data from participants for Internet-based studies that take longer to complete (e.g., Crawford et al., 2001; Fan & Yan, 2010). In fact, meta-analyses have indicated that survey length is one of the strongest contributors to response rates (Edwards et al., 2002). Based on rates of recruitment, study completion, and feedback received from participants, we find that individuals are generally willing to complete 5–10 minute studies without any sizeable monetary recruitment incentive. We have further found that individuals are reasonably willing to complete 15–20 minute studies. However, when studies without sizeable recruitment incentives take longer than 20–30 minutes, recruitment is noticeably more difficult, dropout rates tend to be higher, and participants are more likely to express frustration at the length of the study.

Time-stamp data from thousands of online participants collected in our lab

suggest that individuals can typically complete roughly 10 questions per minute (depending somewhat on item length/complexity, number and complexity of response options, and participant-based factors like reading level, conscientiousness, attention, and individual self-awareness). Thus, response rates may be maximized in Internet-based studies with fewer than 200–300 questions or that take no longer than 20–30 minutes to complete, especially if participants are volunteers or receive only minimal recruitment incentives. It is worth noting that participants seem to be more tolerant of longer studies when they find those studies intrinsically interesting (e.g., there may be greater tolerance for a 30-min survey about sex than one about self-concepts). Intrinsic interest is robustly linked to higher response rates in survey research (e.g., Cook, Heath, & Thompson, 2000; Fan & Yan, 2010), representing one of the most influential factors associated with response rates in meta-analyses (Edwards et al., 2002) .

Managing Dropout

While it is rare for participants to leave during a laboratory study, such dropouts are much more common in Internet-based research (e.g., O’Neil & Penrod, 2001), perhaps because of the greater anonymity and decreased oversight in such studies (Birnbaum, 2004). Dropouts can harm external validity: If dropout differs by demographic or other factors (e.g., conscientiousness, age, employment status), the results may not generalize to groups that are more likely to drop out. Thus, it is essential to examine whether or not dropouts differ systematically on key variables – if so, controlling for these variables in analyses may minimize the bias (Singer & Willett, 2003, p. 158; also see Tabachnick & Fidell, 2007, and Mazza & Enders, Chapter 24 in this volume, for advice on handling missing data). Dropout may also harm internal validity: If dropout differs across experimental conditions (e.g., if one condition is inherently more interesting or rewarding), condition may be confounded with motivation (Birnbaum, 2004). Because of these concerns, Reips (2002) recommends routinely reporting dropout rates (separately for each experimental condition) and relevant analyses in all Internet-based research. When differential dropout across conditions is a major concern, this confound can be minimized by including several pages of questions prior to introducing the experimental manipulation (Reips, 2002). This approach can reduce confounding of motivation across conditions by allowing unmotivated participants to drop out before the manipulation. However, such an approach may further bias the sample and decrease external validity.

Of course, the best approach to managing dropout is to take steps to minimize dropout while designing the study. Many of the design factors already discussed can substantially influence dropout rates. For instance, dropout can be minimized by avoiding technical barriers to participation and maximizing accessibility (e.g., avoiding Java applets and other client-side programming unless essential; using formatting and programming that is compatible with mobile devices), keeping the study as short as possible, offering appropriate incentives, using attractive and professional formatting, and including interesting questions or tasks.

Pretesting

As with lab-based studies, it is important to pilot-test an online study. In addition to ensuring that the study website is working correctly (including custom programming like question branching and random assignment) and that the measures are free of formatting errors, researchers should also double-check to ensure that data are being saved properly. It is also helpful to test a study across different Web browsers and computers to ensure high levels of compatibility. In practice, this may involve asking study team members to complete a newly programmed study several times each and then downloading the resulting data. Finally, we strongly recommend asking participants for open-ended comments at the end of the study. Feedback may provide an early warning for identifying technical problems, and it shows participants greater respect, offering frustrated participants a place to express their concerns.

Data Cleaning

Given that the basic design of studies conducted on the Internet often does not differ dramatically from studies conducted in more traditional environments, data cleaning in Internet-based studies is typically very similar. However, there are a few considerations that are particularly relevant to Internet-based data.

Screening for Multiple Submissions. For numerous reasons, Internet-based studies may be more prone to multiple submissions than are other methods (Johnson, 2005). A common cause of multiple submissions stems from participants hitting the submit button more than once, unintentionally submitting multiple copies of the same data. This can happen when a server lag or a browser glitch makes it appear as if nothing has happened immediately after clicking the submit button. These duplicate submissions are easily caught by sorting the responses by date of submission and looking for identical responses submitted in

very quick succession. Johnson (2005) found that 3.8% of the responses to an online survey represented such duplications.

Multiple submissions may also occur when a respondent cannot complete the study in a single session, closes the Web browser, and returns to the webpage again later to complete the entire study from the beginning, generating two rows of data (one complete and one incomplete). In such cases, it can help simply to ask participants if they have completed the study before. In studies offering generous incentives, individuals sometimes participate multiple times in an attempt to receive additional payments or lottery entries. Such redundant submissions can sometimes be caught using identifying information saved in the survey (e.g., email addresses, ID numbers embedded in follow-up survey links). Multiple submissions can also be identified by looking for rows of data with identical IP addresses and then comparing responses on demographic questions. Although IP addresses can be helpful in screening for duplicate submissions, IP addresses are not completely reliable identifiers, as different individuals in the same household or even the same university may appear to have the same IP address. Furthermore, some Internet service providers may assign the same IP address to different computers at different times. Therefore, IP addresses should be used only in conjunction with other sources of information when identifying duplicates. Luckily, duplicate submissions are rare and likely do not affect results much in large samples (Birnbaum, 2004).

Screening for Inattentive and Invalid Responses. Although research has not found that Internet-based data is inherently less valid than data collected using other methods, we believe that all researchers should routinely screen for evidence of excessive inattention or invalidity, and Internet-based studies offer some unique methods of doing this. As discussed previously, researchers may use measures that identify highly inattentive responses (such as the ARS; Maniaci & Rogge, 2013a, or one of the indicators described by Meade & Craig, 2012) to eliminate this potential source of error variance. Inattention can also be gauged by simply tracking the length of time it takes each participant to complete a study or portions of a study, as impossibly low study completion times serve as an indicator of inattention (Meade & Craig, 2012). It may also be important in some studies to ensure that participants complete the study in a single sitting or within a specific time frame rather than on multiple occasions (e.g., experiments involving priming or mood manipulations). Although some study platforms include the date and time the study was started and completed by default, such time stamps can also be recorded using JavaScript code. For studies involving audio/visual clips as part of a manipulation, researchers can

assess compliance by using JavaScript to record the amount of time spent watching the clip. Internet-based studies can also include direct manipulation checks where participants are asked to relate instructions back to the researcher via an open-ended text response or describe what they viewed in a video clip. As the researcher is not present to monitor compliance as in a laboratory study, such direct and indirect measures of compliance and attention become more critical with Internet-based studies.

Screening for Outliers. Statisticians have long called for researchers to examine their data sets for outliers, as outliers can radically skew results (e.g., Tabachnick & Fidell, 2007; McClelland, Chapter 23 in this volume). Social psychologists often ignore this issue, perhaps assuming that replicating an effect across several small samples is sufficient to reduce the risk of spurious findings resulting from unexamined outliers. We strongly recommend screening for multivariate outliers as it not only helps address this statistical concern, but it can also help identify a different type of inattentive response (e.g., consistent yet exaggerated/nonsensical responses generated by individuals not taking the study seriously). Although researchers are sometimes loathe to eliminate responses (one of the possible methods of dealing with outliers), the large samples afforded by Internet-based studies alleviate this concern. For more discussion of identifying and dealing with statistical outliers and other statistical assumptions, see Tabachnick and Fidell's (2007) excellent chapter on data cleaning and McClelland's (Chapter 23 in this volume) contribution on dealing with “nasty” data.

Recruitment and Sampling Strategies

Implementing a study online need not change how a researcher recruits participants, as traditional recruitment methods (e.g., psychology participant pool, flyers, newspaper and radio ads, random-digit dialing) can still be used to obtain a sample. However, implementing a study on the Internet affords several unique recruitment options. Often, multiple recruitment methods are combined. In such cases, recruitment source can be monitored for later comparison by adding code to the link (e.g., “...index.html?source=forum”) that can be saved by the study webpage.

Online Forums and Websites

As the Internet access has broadened over the past two decades, online

communities have developed where individuals struggling with specific issues (e.g., parenting problems, specific physical or mental illnesses) or sharing a common interest (e.g., collectors, hobbyists, gaming enthusiasts) interact with one another on an ongoing basis. These online communities can represent a gold mine of recruitment resources because they enable savvy researchers to recruit from target populations simply by posting in online forums. Such communities can provide unprecedented access to rare populations (e.g., as mentioned previously, Meier *et al.* 2011 recruited a sample of 400 female-to-male transsexuals by advertising on websites and discussion forums dedicated to this group). Some researchers have also worked with websites to achieve more direct access to participants. For instance, Gebauer, Leary, and Neberich (2012) obtained access to data from more than 11,000 users of a German online dating website, finding that individuals with less popular first names were more likely to be neglected by potential partners.

Several specific strategies may help researchers recruit participants from such online communities and websites more effectively.

- *Pick relevant communities.* Response rates are higher when individuals perceive greater personal relevance or interest in the topic of the study (e.g., Edwards, 2002). Thus, posting a study regarding political attitudes on an online political forum is likely to be far more effective at recruiting participants (and far less likely to trigger irritation within that community) than posting the same study on a forum within an online gaming website.
- *Pick one relevant forum at a time on each site.* Web communities usually have many forums, each focused on a specific topic within that community. It is considered rude (and may get a researcher kicked off a site) to post the same message in more than one forum at a time, as that may be considered “spamming.” Therefore, it is most effective to post to the single most relevant forum on each site and, if necessary, delete that post and start a new message once traffic has stopped.
- *Actively join each community.* Many online communities take their privacy seriously and are disinclined to allow relative strangers to post on their site, particularly if they perceive that posting to be spam. To prevent this impression, a researcher could join a site and participate in the forum by posting replies to other topics. If the forum focuses on the researcher's topic area, he or she may post simple advice and findings from the relevant empirical literature. Such posts are often welcomed by online communities (when tempered with respect and tact) and go a long way to ensure that

researchers will not be viewed as opportunistic interlopers when they post their own studies.

- *Personalize your message.* We recommend that researchers personalize their pleas for participants. Explaining that they are professors or students working on a research project at a specific university can lend credibility and relevance to a project, increasing response rates. Explaining how the research is relevant to that community (and perhaps offering to share a summary of findings after the study is complete) can help build goodwill.
- *Show respect.* Individuals can become very invested in online communities, particularly those that serve as informal support groups. As a result, researchers should show an appropriate level of respect in all posts and email exchanges. If the researcher has any doubt about the members of a site being comfortable with study recruitment posts, the researcher should contact the forum moderator to ask for permission before posting.
- *Promptly respond to all posts.* Forums are meant to be interactive, and so members of the site will naturally reply to a post by a researcher. Most of these posts will be fairly neutral or positive. However, invariably some replies will be negative or hostile. Replying promptly to all posts (positive and negative) with professionalism, respect, and kindness can help sustain the relevance of the thread, foster goodwill, and acknowledge concerns.

Websites Listing Online Studies

Another recruitment option involves posting on search engines so that individuals searching for relevant topics can find the study website. This may be facilitated by creating a carefully designed initial webpage that briefly describes the study in a simple, attractive format before presenting more detailed information. In addition to relying on search engines, researchers can recruit using websites that list Internet-based social science studies. At the time of writing, the two most popular examples for social-personality psychology are a site maintained by John Krantz from the psychology department of Hanover College (<http://psych.hanover.edu/research/exponnet.html>) and a site hosted by the Social Psychology Network and maintained by Scott Plous at Wesleyan University (<http://www.socialpsychology.org/expts.htm>). Both of these sites are free for researchers. As both sites simply require an email exchange to get a study listed, the return on investment is quite high even if the rate of recruitment is not as rapid as it is with other methods. However, recruiting from such sources may introduce additional selection bias, as subjects would have sought out listings of research studies, or specifically searched for relevant keywords.

Online Advertising

Researchers may also recruit participants using online advertising. A popular free option is Craigslist.org, which offers free online classified advertisements for various cities in the United States. Other online advertising services typically have two different pricing options: paying for the number of people who see an ad (impressions) or paying for the number of people who click on an ad (clicks). A researcher has control over cost within either pricing system and can set daily limits, with the caveat that lower payments mean that fewer people will see the ad. For example, Google AdWords allows individuals to design short text ads that will appear as sponsored links following Google searches. Researchers can post an ad and select keywords or phrases that will trigger their ads (e.g., “prejudice test” for a study on implicit prejudice). Google AdWords tracks the relative popularity and success of each keyword in an ad campaign so that advertisers can quickly optimize the terms being used as well as the price they are willing to pay for impressions or clicks.

Similarly, social networking services like Facebook sell advertising that is displayed when viewing Facebook profiles. One marked advantage of Facebook advertising is that advertisers can target their ads to specific demographics (e.g., only showing the ad to women), narrowing recruitment based on age, gender, relationship status, and geographic location. Although online advertising tends to be less expensive in cost per response than many forms of traditional paid advertising (e.g., print ads, radio spots, mailing postcards or flyers), we have noted a marked rise in the cost of paid online advertising over the last decade.

Email Distribution Lists and Listservs

Another recruitment option involves contacting the administrators of email distribution lists to request that they send a mass email advertising a study to their members. Many large organizations routinely collect email addresses as a means of quickly contacting members and may also set up listservs (special email services that allow members to contact the entire group directly). Listservs are essentially email-based forums in which the members of a group hold ongoing discussions. Email distribution lists and listservs can therefore function as a powerful method of recruitment in a manner similar to posting on web forums, allowing researchers to recruit targeted audiences. As a result, we offer similar advice to that offered for online forums: pick highly relevant listservs, personalize your message, emphasize your academic affiliation if possible, show respect for the group members, and promptly respond to all replies. Response

rates can be markedly enhanced if the email comes directly from someone in a position of leadership in the group (e.g., the president or CEO of an organization, the founder of a website or group). Thus, researchers interested in using email distribution lists or listservs as a method of recruitment might start by personally approaching the administration of the corresponding groups in an effort to secure their support. Researchers should also check listserv policies to see if they allow recruitment messages (e.g., the APA listserv currently prohibits such messages, although the SPSP listserv allows them).

Enhancing Response Rates from Email Invitations

Methodological research has robustly demonstrated that repeated invitations yield higher response rates than single invitations (e.g., Cook et al., 2000; Dillman, 2007; Edwards et al., 2002; Fan & Yan, 2010). Other research suggests that sending pre-notifications (e.g., a text message or postcard notifying potential participants to expect an invitation email; Bosnjak, Neubarth, Couper, Bandilla, & Kaczmire, 2008) and personalizing those invitations (e.g., adding the researchers' names, job titles, affiliations, and digital signatures; Cook et al., 2000) can enhance response rates. Mentioning some form of scarcity (e.g., potential participants are part of a small selected group, the participation deadline is approaching) can further boost response rates (Edwards et al., 2002). With email invitations, researchers should carefully craft messages to avoid spam filters. This means not using attachments, sending individual emails (rather than adding a large number of email addresses to each email in an effort to save time), avoiding the use of all capital letters and words commonly used by bulk mailers (e.g., "free," "win," "click here!!"), including full contact information (e.g., a phone number and mailing address), and using each respondent's name, if possible. Email invitations should also come from a university email address to emphasize academic affiliation.

Probability-Based Internet Panels

Despite concerns that Internet-based samples are not representative, some researchers have used traditional approaches (e.g., random-digit dialing) to form probability-based Internet panels that are representative of the larger population. For example, Knowledge Networks (www.knowledgenetworks.com) created a nationally representative Internet panel of households in the United States by using random-digit dialing and U.S. Postal Service records of residential households. Households in the panel that did not initially have Internet access

were sent laptop computers along with an Internet connection paid for by Knowledge Networks. In addition to purchasing access to the Knowledge Networks panel, researchers can also conduct experiments using this representative panel for free (albeit with stringent limitations on length and content) by submitting a proposal to Time-Sharing Experiments for the Social Sciences (TESS; <http://www.tessexperiments.org>, a program that was originally supported by the National Science Foundation). Researchers outside the United States have similarly arranged Internet panels representative of the population in other countries. For instance, the Longitudinal Internet Studies for the Social Sciences (LISS; <http://www.lissdata.nl>) panel is a probability-based Internet panel of about 5,000 households in the Netherlands in which panel members participate in studies each month in exchange for financial incentives and Internet access (Das, 2012; Scherpenzeel & Das, 2011). Like TESS, the LISS panel is available for free to researchers whose proposals are approved, including those outside of the Netherlands. Although representative panels are much more difficult to create and maintain than are traditional Internet samples, they offer a solution to researchers seeking a truly representative sample.

Crowdsourcing

A more recent recruitment innovation is the use of crowdsourcing – services that offer small financial incentives in return for Internet users completing relatively short and straightforward tasks. One popular example is Amazon.com's Mechanical Turk (MTurk; www.mturk.com), which boasts access to more than 500,000 registered users from 190 countries. Individuals can log into their MTurk accounts, select and complete jobs (called “Human Intelligence Tasks” or HITs), and then (upon approval from the poster of each HIT) receive payments directly to their Amazon.com store accounts. A researcher can simply create an MTurk HIT with a link to their Internet-based study that instructs workers to follow the link, complete the study, and then return to the MTurk page to type in a codeword placed on the last page of the study by the researcher (as evidence that they completed the study). MTurk also allows HITs to be limited based on geographic location and the quality of a user's previous responses (e.g., the proportion of completed HITs that were approved by the poster). Although MTurk is one of the most popular crowdsourcing services, there are other similar services. For instance, CrowdFlower (crowdflower.com) claims to have access to more than 2.5 million users from more than 50 distinct partner sources, including MTurk users.

Because the crowdsourcing service provider (e.g., Amazon.com in the case of MTurk) handles the monetary transactions, such services provide a robust infrastructure for researchers to easily offer online respondents small payments that would not be practical by other means, such as disbursement checks. Researchers also do not need to collect personal information, allowing participants to remain anonymous. We have used MTurk to recruit more than 300 people for a 10-minute survey in 1–2 weeks for \$33 (10 cents per respondent plus the 10% processing fee charged by Amazon.com). As a rule of thumb, we find that paying participants 1 cent per minute of a study (e.g., 20 cents for a 20-minute study) is about the minimum payment that will yield a reasonable response rate (e.g., 10–20 people per day). However, it is easy to adjust this payment to achieve faster recruitment rates. Buhrmester, Kwang, and Gosling (2011) demonstrated that different payment amounts (ranging from 2 to 50 cents) affected the rate of recruitment from MTurk, but did not affect the quality of data obtained.

Although crowdsourcing activities are typically simple tasks that individuals can complete quickly, crowdsourcing can also be used for more extensive studies. For instance, by asking participants recruited from MTurk to enter their MTurk identification number into a survey (or, alternatively, providing them with a unique, randomly generated ID number on the last page of the survey and asking them to type that number into the MTurk page), it is possible to offer monetary bonuses for complying with additional study elements (e.g., completing longitudinal follow-up assessments). By using MTurk identification numbers, researchers can contact participants directly through the MTurk interface (e.g., to invite them to complete a follow-up survey) without obtaining identifying information.

Because crowdsourcing involves small incentives, initially there were concerns that respondents might be particularly inattentive. However, studies comparing MTurk samples to other recruitment sources show comparable data quality (e.g., internal consistency and measurement invariance of scales; Behrend, Sharek, Meade, & Wiebe, 2011) and patterns of findings across those samples. Paolacci, Chandler, and Ipeirotis (2010) compared data collected from MTurk, Internet discussion board postings, and an undergraduate participant pool, finding that several experimental manipulations replicated previous findings from the judgment and decision-making literature in all three groups. Furthermore, participants from the three samples did not differ in attention, as assessed using the question, “While watching the television, have you ever had a

fatal heart attack?” In our own work, we have compared data collected from MTurk samples to data from (1) other online participants, (2) students completing the study online, and (3) students completing the same study in our lab supervised by an experimenter (Maniaci & Rogge, 2013b). These groups exhibited similar compliance with study tasks (e.g., watching a video clip) and levels of attention on various indicators, provided data with comparable internal consistencies, and yielded similar patterns of results. Thus a growing body of literature suggests that MTurk respondents are sufficiently attentive and compliant.

There were also initial concerns that samples obtained via crowdsourcing might be less demographically representative than samples obtained by other methods. Although MTurk samples are certainly not representative of the population, they do offer considerable demographic diversity – perhaps more so than typical Internet samples. For instance, Buhrmester *et al.* (2011) reported that their sample of MTurk participants included individuals from all 50 states in the United States and more than 50 other countries,⁶ and that compared to a standard Internet sample, the MTurk sample was slightly older and included a greater proportion of non-white participants. Behrend *et al.* (2011) also found that MTurk participants were older, had more work experience, and were more ethnically diverse than an undergraduate sample. Thus, when compared to the relative homogeneity of undergraduate student samples, MTurk samples continue to demonstrate the advantages of increased demographic diversity offered by online samples (e.g., greater variance on age, education, income, marital status). Another concern about MTurk recruitment is that participants can easily search for and participate in multiple studies posted by a specific researcher, reducing the independence of samples. However, the online participant pool is by definition orders of magnitude larger than commonly used undergraduate research pools. Furthermore, if repeat participation is a concern, it is possible in MTurk to restrict participation to individuals who have not participated in previous studies.

Snowball Recruitment

A final recruitment option particularly amenable to Internet-based research is snowball recruiting via email or social networking services, which involves encouraging participants to contact family and friends to elicit their participation. This approach is useful for sampling rare or hard-to-reach populations. For instance, Simon and Ruhs (2008) studied the effects of dual

identity on political activity by using snowball sampling to recruit Turkish migrants living in Germany. This strategy can be particularly effective with Internet-based studies as individuals can post links to a study on social networking sites (e.g., Facebook, Twitter, personal blogs) or email the link to a large number of friends with little effort; indeed, participants sometimes do this spontaneously if they enjoyed participating in the study. It is also possible to add JavaScript code at the end of a study to allow respondents to share the study on social networking sites by simply clicking a button. Snowball recruitment tends to be most effective when individuals are motivated to invest time and effort in sharing the study, which is most likely when studies offer useful feedback or other incentives. However, as snowball recruiting draws on interdependent social networks, its use may reduce the representativeness of the resulting sample and could introduce unmeasured dependencies in the data.

Monetary Recruitment Incentives

Much research examining the relative efficacy of different recruitment incentives for social science research suggests that both monetary and nonmonetary (e.g., small gifts) incentives tend to increase response rates (e.g., Edwards et al., 2002) and decrease dropout rates (Görizt, 2006). With Internet-based studies, it is easy to send small payments using crowdsourcing services or to send electronic gift certificates via email. Meta-analyses including 58 Internet-based experiments concluded that material incentives significantly increased response rates (the proportion of invited participants who access the study website) by an average of 19% and retention rates (the proportion of respondents who complete the entire study) by an average of 27% (Görizt, 2006). Furthermore, lottery incentives (i.e., raffles with prizes ranging from approximately \$50 to \$200) were no less effective than other types of incentives, regardless of the amount of the lottery. Although larger incentives are generally more effective, size of monetary incentives is not linearly related to response rates – that is, very large monetary incentives tend to yield diminishing returns (e.g., Fan & Yan, 2010). This is consistent with our experience with MTurk, where even nominal payments of 10–30 cents are highly effective.

Feedback as a Recruitment Incentive

In her meta-analysis, Görizt (2006) found that offering participants a summary of the research findings as a recruitment incentive was no less effective than other types of incentives, including lotteries, gifts, and monetary payment. We

have also found that providing individualized feedback at the end of a study based on each participant's responses can be a highly effective recruitment incentive. The feedback can be as simple as calculating total scores on a scale and then presenting a short paragraph describing what that score means. It may also include text corresponding to the specific range of scores within which the respondent falls (e.g., high, medium, and low). Adding graphics showing an individual's score relative to the larger distribution (see [Figure 17.1](#)) can enhance the perceived value of feedback. We often give feedback on 10 or more separate dimensions at the end of a study, providing respondents with a detailed description of how they compare to thousands of other respondents. Although programming such feedback requires additional work while preparing the study, it offers a high return on investment by enhancing recruitment for the duration of the study. It may not be appropriate to provide feedback in certain studies – for instance, researchers may wish to avoid providing detailed feedback on key outcomes in the beginning of a longitudinal study so as to preclude influencing later responses. In other cases, providing potentially negative or threatening feedback (e.g., regarding one's intelligence or standing on other socially desirable characteristics) could induce distress. When feedback might be potentially surprising or distressing (e.g., a measure of implicit prejudice), researchers should word feedback carefully and avoid overstating the meaning of any one score.

In addition to offering feedback for specific studies, some researchers have created websites that offer feedback on multiple personality measures, which may be updated to accommodate new research. One such site is www.outofservice.com, which has collected personality assessments from nearly 9 million respondents since 1997, contributing to more than 15 empirical articles. For instance, Soto *et al.* (2011) used this site to examine age differences in personality traits using a sample of more than 1.2 million respondents. Rentfrow, Gosling, and Potter (2008) used data from the site to explore geographic differences in personality traits across all 50 states in the United States. Similarly, the Project Implicit website at implicit.harvard.edu has collected millions of IAT responses that have been used in numerous studies since the site launched in 1998, including one paper that analyzed 2.5 million IATs completed over a period of 6 years (Nosek et al., 2007). Both of these sites offer individualized feedback as their primary recruitment incentive.

Ethical Issues

Along with enabling efficient recruitment of large and diverse samples, Internet data collection presents a variety of unique ethical concerns.

Informed Consent

A large portion of social-personality research can be considered of such minimal risk to participants that it is typically exempt from IRB review and does not require documentation of informed consent (e.g., surveys that are anonymous or ask relatively benign questions). Although rules and standards vary across institutions, participants in such studies can usually simply be given an information page describing the nature of the study and the risks and benefits of participation. If a study is not exempt, many IRBs may still provide a waiver of documentation of consent if the study is of no more than minimal risk and involves only activities that would not require written consent outside of a research context. With such a waiver, a signed consent form is not required and participants may merely read a description of the study and perhaps click a button to signify that they freely choose to participate.

Studies that involve greater than minimal risk demand a more thorough process of informed consent, which may involve (1) direct interaction with participants and (2) a signed informed consent form. Such studies can be more challenging to implement on the Internet as making informed consent an interactive process could necessitate direct contact with potential participants via email or phone. Some IRBs will also not accept electronic signatures in place of a physical, signed copy of the consent form, and the legality of electronic signatures may vary across participants' jurisdictions. In such cases, potential participants may need to mail a signed consent form to the researchers and wait to receive a link to the study. This process places greater barriers to participation on potential participants (in contrast to minimal risk studies where participants can begin immediately) and may lower response rates considerably. Thus, implementing a greater than minimal risk study on the Internet might incur significant additional costs to achieve a thorough informed consent process.

Public vs. Private Behavior

One challenge facing researchers and ethics review boards is determining whether certain types of Internet-based communication are considered public behavior (and therefore open to study without obtaining informed consent) or private behavior (requiring more direct consent). Kraut *et al.* (2004) argued that

communication in online forums with unrestricted membership (e.g., public Internet forums) should be treated as public behavior, as posters do not have a reasonable expectation of privacy in such settings. However, Kraut *et al.* (2004) suggested that the decision of whether there is a reasonable expectation of privacy must be made on a case-by-case basis, taking into account the nature of a specific forum and its membership. For instance, Facebook users who choose the highest levels of privacy settings for their accounts likely have a greater expectation of privacy than users choosing lower privacy settings. Researchers must be sensitive to these issues in order to maintain respect for participants' right to privacy.

Maintaining Data Security

As noted previously, any information transmitted over the Internet involves some risk of interception by unintended third parties. Although this risk is also present in studies using more traditional methods (e.g., phone surveys could be overheard, physical data could be misplaced, data stored on computers connected to the Internet could be intercepted by computer hackers), the risk may be greater with information transmitted over the Internet. One response to this risk is to avoid collecting identifying information or to save identifying information separately from other responses (using a code to link the two). Regardless of how a study is conducted, ensuring that data stored electronically are fully de-identified is the safest method of protecting confidentiality. When data include both identifying information and responses to questions that could cause harm or embarrassment if disclosed (e.g., illegal activity, infidelity), then researchers should take steps to ensure that survey responses will be encrypted before transmission (e.g., using SSL or another form of HTTPS encryption). If employing a study-hosting service, researchers may want to inquire about the service's security practices, both in terms of data storage and transfer and physical security of the Web servers. Some IRBs may express concerns about storing data on the servers of a study-hosting site, where it could be accessible to individuals not in the research team. Even after a researcher downloads and deletes responses from the study-hosting site, the site may retain backups or logs of participants' IP addresses and responses. If this is a concern, then hosting the study on a university server or one's own server may provide greater control over data storage and security. Researchers should also use common sense when using Web-based services. Even if a service employs advanced encryption and security procedures, there could still be a breach of confidentiality if the researcher's password is discovered (or provided to research assistants or

colleagues) or if data are transmitted or stored insecurely (e.g., sending data files by email).

IP Addresses and Anonymity

Many study-hosting services automatically save IP addresses with the data for each participant. Although those IP addresses can be useful in tracking duplicate submissions or in pairing up longitudinal responses, they can also be considered a form of identifying information, in that they can provide geographical information (city, state, zip code) for the computer used to complete the study. Thus, to truly de-identify a data set, it is necessary to delete IP addresses in addition to other identifying information. Some study hosting services allow researchers to disable the automatic saving of IP addresses, allowing for truly anonymous data collection.

Greater than Minimal Risk Studies

Although most Internet-based studies involve minimal risk to participants (e.g., answering benign questions or engaging in a brief experimental task), it is possible to conduct research over the Internet with greater than minimal risk. For instance, an Internet-based study may ask about illegal behavior or include experimental manipulations with the potential to cause psychological harm (e.g., providing false feedback with negative implications for self-esteem). Although these risks are not inherently different from research conducted in a laboratory setting, Internet-based research may reduce the researcher's ability to monitor participants' reactions and to respond appropriately (Kraut et al., 2004). If a participant reacts unusually negatively to a manipulation in a laboratory setting, the experimenter could stop the experiment, debrief the participant, and refer them to appropriate resources. With Internet-based studies, researchers typically cannot directly observe such aversive reactions or respond to them as immediately. Nevertheless, Barchard and Williams (2008) argue that researchers can ethically conduct greater than minimal-risk Internet-based studies as long as they take additional steps to ensure that participants understand the risks (e.g., assessing comprehension when obtaining informed consent) and to monitor and respond to potential adverse reactions (e.g., collecting contact information to follow up with distressed participants; making the study website available only at times when participants can contact the researcher directly via phone, email, or a chat room).

Deception and Debriefing

Further complications arise in Internet-based studies involving deception. Ethical guidelines require a formal process of debriefing in which the researcher explains the deception and why it was necessary, and takes steps to correct participants' misconceptions (Smith, Chapter 3 in this volume). Ideally, this debriefing process is interactive, providing participants with ample opportunity to have any questions or concerns addressed and allowing the researcher to carefully probe for suspicion. Although it is possible to implement an automated debriefing in an Internet-based study, this format limits the degree to which debriefing can be interactive. If deception is relatively minor (e.g., failing to fully explain the true purpose of a study), then it may be appropriate to simply include debriefing text as the last webpage in a study. However, because Internet-based studies tend to have higher dropout rates than do laboratory studies (O'Neil & Penrod, 2001), researchers may need to take steps to ensure that all participants will be debriefed (e.g., emailing participants or using a pop-up window to provide information if a participant prematurely exits the study; Nosek, Banaji, & Greenwald, 2002). If the deception entails greater than minimal risk of harm to participants, Barchard and Williams (2008) recommend telling participants that they will be fully debriefed after the study, asking questions to ensure that participants understand the debriefing information, and following up with an email as needed. Because of these complications, some researchers (e.g., Birnbaum, 2004; Fraley, 2004) argue that deception is difficult to justify in Internet-based research.

Concluding Comments: The Future of Internet-Based Research

With the rapid expansion of the Internet into the daily lives of individuals worldwide, social-personality psychologists have increasingly utilized the Internet both as a new source of implementing traditional study designs and as an opportunity to develop unique designs that capitalize on developing technology. The Internet offers clear benefits over traditional methods, allowing researchers to collect large, adequately powered, and relatively diverse samples with minimal costs, reach populations that would otherwise be difficult to study, utilize a broad array of designs, and automate time-consuming and error-prone research tasks. The Internet will likely continue to penetrate into broader portions of society in the next decade, and new technology will enable

increasingly innovative research designs. For instance, 54% of the U.S. population currently has access to the Internet on a mobile device, and this proportion has increased steadily over the past 5 years (International Telecommunication Union, 2012). Technological advances are likely to enable the collection of increasingly sophisticated types of data using the Internet (e.g., examining patterns of social proximity and interaction using wireless technology, using GPS and accelerometers to monitor activity and energy expenditure using Internet-enabled mobile devices, and remotely collecting ambulatory physiological data). As mobile Internet access and smartphone/tablet technology advance, it may soon be possible to utilize the Internet to recruit large, diverse samples that can provide rich longitudinal, experience sampling, and even physiological data in participants' natural environments.

The Internet holds great promise for advancing research and theory while addressing some of the problems that have long plagued social-personality psychology (e.g., overreliance on student samples with limited diversity, underpowered studies). At the same time, the Internet presents challenges in screening participants, monitoring attention and compliance, implementing certain designs, addressing ethical concerns, and obtaining representative samples. Accordingly, we urge researchers to view the Internet as a promising tool that complements more traditional methods (e.g., laboratory studies, behavioral observation). While researchers should be aware of both the pros and cons of collecting data on the Internet, they would be remiss if they failed to capitalize on the promise of this advancing technology.

References

- Amichai-Hamburger, Y. (2005). Internet minimal group paradigm. *CyberPsychology & Behavior*, 8, 140–142.
- Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63, 596–612.
- Avendano, M., Scherpenzeel, A. C., & Mackenbach, J. P. (2011). Can biomarkers be collected in an Internet survey? A pilot study in the LISS panel. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the Internet: Advances in applied methods and research strategies* (pp. 371–412). New York: Routledge.
- Bagozzi, R. P., Dholakia, U. M., & Mookerjee, A. (2006). Individual and group

- bases of social influence in online environments, *Media Psychology*, 8, 95–126.
- Barchard, K. A. & Williams, J. (2008). Practical advice for conducting ethical online experiments and questionnaires for United States psychologists. *Behavior Research Methods*, 4, 1111–1128.
- Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements. *Perspectives on Psychological Science*, 2, 396–403.
- Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavioral Research Methods*, 43, 800–813.
- Binning, K. R., & Sherman, D. K. (2011). Categorization and communication in the face of prejudice: When describing perceptions changes what is perceived, *Journal of Personality and Social Psychology*, 101, 321–336.
- Birnbaum, M. H. (2004). Methodological and ethical issues in conducting social psychology research via the Internet. In C. Sansone, C. C. Morf, & A. T. Panter (Eds.), *Handbook of methods in social psychology* (pp. 359–382). Thousand Oaks, CA: Sage.
- Bogart, K. R., & Matsumoto, D. (2010). Facial mimicry is not necessary to recognize emotion: Facial expression recognition by people with Moebius syndrome. *Social Neuroscience*, 5, 241–251.
- Bosnjak, M., Neubarth, W., Couper, M. P., Bandilla, W., & Kaczmire, L. (2008). Prenotification in Web-based access panel surveys – The influence of mobile text messaging versus e-mail on response rates and sample composition. *Social Science Computer Review*, 26, 213–223.
- Brock, R. L., Barry, R. A., Lawrennce, E., Dey, J., & Rolffs, J. (2012). Internet administration of paper-and-pencil questionnaires used in couple research: Assessing psychometric equivalence. *Assessment*, 19, 226–242.
- Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, 90, 125–144.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on*

Psychological Science, 6, 3–5.

- Burke, B. L., Martens, A., & Faucher, E. H. (2010). Two decades of Terror Management Theory: A meta-analysis of mortality salience research. *Personality and Social Psychology Review*, 14, 155–195.
- Clark-Carter, D. (1997). The account taken of statistical power in research published in the British Journal of Psychology. *British Journal of Psychology*, 88, 71–83.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145–153.
- Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in Web-or Internet-based surveys. *Educational and Psychological Measurement*, 60, 821–836.
- Couper, M. P., Conrad, F. G., & Tourangeau, R. (2007). Visual context effects in web surveys. *Public Opinion Quarterly*, 71, 623–634.
- Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001). Web surveys: Perceptions of burden. *Social Science Computer Review*, 19, 146–162.
- Das, M. (2012). Innovation in online data collection for scientific research: The Dutch MESS project. *Methodological Innovations Online*, 7, 7–24.
- Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design method. 2007 update with new Internet, visual, and mixed-mode guide* (2nd ed.). New York: John Wiley & Sons.
- Edwards, P., Roberts, I., Clarke, M., DiGuseppi, C., Prata, S., Wentz, R., & Kwan, I. (2002). Increasing response rates to postal questionnaires: Systematic review. *British Medical Journal*, 324, 1183–1192.
- Erdle, S., Gosling, S. D., & Potter, J. (2009). Does self-esteem account for the higher-order factors of the Big Five? *Journal of Research in Personality*, 43, 921–922.
- Etter, J. F., Neidhart, E., Bertrand, S., Malafosse, A., & Bertrand, D. (2005). Collecting saliva by mail for genetic and cotinine analyses in participants recruited through the Internet. *European Journal of Epidemiology*, 20, 833–838.

- Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior*, 26, 132–139.
- Fraley, R. C. (2004). *How to conduct behavioral research over the Internet: A beginner's guide to HTML and CGI/Perl*. New York: Guilford.
- Frost, J. H., Chance, Z., Norton, M. I., & Ariely, D. (2008). People are experience goods: Improving online dating with virtual dates. *Journal of Interactive Marketing*, 22, 51–61.
- Funk, J. L., & Rogge, R. D. (2007). Testing the ruler with item response theory: Increasing precision of measurement for relationship satisfaction with the Couples Satisfaction Index. *Journal of Family Psychology*, 21, 572–583.
- Gebauer, J. E., Leary, M. R., & Neberich, W. (2012). Unfortunate first names: Effects of name-based relational devaluation and interpersonal neglect. *Social Psychological and Personality Science*, 3, 590–596.
- Glaser, J., Dixit, J., & Green, D. P. (2002). Studying hate crime with the Internet: What makes racists advocate racial violence? *Journal of Social Issues*, 58, 177–193.
- Göriz, A. S. (2006). Incentives in Web studies: Methodological issues and a review. *International Journal of Internet Science*, 1, 58–70.
- Göriz, A. S. (2007). The induction of mood via the WWW. *Motivation and Emotion*, 31, 35–47.
- Göriz, A. S., & Crutzen, R. (2012). Reminders in web-based data collection: Increasing response at the price of retention? *American Journal of Evaluation*, 33, 240–250.
- Gosling, S. D., & Johnson, J. A. (Eds.). (2010). *Advanced methods for conducting online behavioral research*. Washington, DC: American Psychological Association.
- Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the Internet in reaching more diverse samples. *Behavioral and Brain Sciences*, 33, 94–95.
- Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies: A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, 59, 93–104.

- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- Guadagno, R. E., & Cialdini, R. B. (2002). Online persuasion: An examination of gender differences in computer-mediated interpersonal influence. *Group Dynamics: Theory, Research, and Practice*, 6, 38–51.
- Hsu, L. M. (1989). Random sampling, randomization, and equivalence of contrasted groups in psychotherapy outcome research. *Journal of Consulting and Clinical Psychology*, 57, 131–137.
- Ijzerman, H., & Semin, G. R. (2009). The thermometer of social relations: Mapping social proximity on temperature. *Psychological Science*, 20, 1214–1220.
- International Telecommunication Union (2012). *Key global telecom indicators for the world telecommunication service sector*. Retrieved March 27, 2012, from http://www.itu.int/ITU-D/ict/statistics/at_glance/KeyTelecom.html.
- Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, 39, 103–129.
- Koo, M., Algoe, S. B., Wilson, T. D., & Gilbert, D. T. (2008). It's a wonderful life: Mentally subtracting positive events improves people's affective states, contrary to their affective forecasts, *Journal of Personality and Social Psychology*, 95, 1217–1224.
- Krantz, J. H., & Dalal, R. (2000). Validity of web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 35–60). San Diego, CA: Academic Press.
- Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of Board of Scientific Affairs' advisory group on the conduct of research on the Internet. *American Psychologist*, 59, 105–117.
- Kruger, J., Epley, N., Parker, J., & Ng, Z. W. (2005). Egocentrism over e-mail: Can we communicate as well as we think? *Journal of Personality and Social Psychology*, 89, 925–936.
- Le, B., Moss, W. B., & Mashek, D. (2007). Assessing relationship closeness

- online: Moving from an interval-scaled to a continuous measure of including others in the self. *Social Science Computer Review*, 25, 405–409.
- Lee, S., Rogge, R. D., & Reis, H. T. (2010). Assessing the seeds of relationship decay: Using implicit evaluations to detect the early stages of disillusionment. *Psychological Science*, 21, 857–864.
- Lenhart, A., Horrigan, J., Rainie, L., Allen, K., Boyce, A., Madden, M., & O’Grady, E. (2003). *The ever-shifting Internet population: A new look at Internet access and the digital divide*. Washington, DC: Pew Internet and American Life Project.
- Maniaci, M. R., & Rogge, R. D. (2013a). *Caring about carelessness: Participant inattention and its effects on research*. Manuscript submitted for publication.
- Maniaci, M. R., & Rogge, R. D. (2013b). *Comparing data quality from laboratory, Mechanical Turk, and other online samples*. Manuscript in preparation.
- Marcus, B., Machilek, F., & Schütz, A. (2006). Personality in cyberspace: Personal web sites as media for personality expressions and impressions. *Journal of Personality and Social Psychology*, 90, 1014–1031.
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17, 437–455.
- Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are Internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study. *Organizational Research Methods*, 10, 322–345.
- Meier, S. C., Fitzgerald, K. M., Pardo, S. T., & Babcock, J. (2011). The effects of hormonal gender affirmation treatment on mental health in female-to-male transsexuals. *Journal of Gay & Lesbian Mental Health*, 15, 281–299.
- Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, 7, 221–237.
- Murray, D. M., & Fisher, J. D. (2002). The Internet: A virtually untapped tool for research. *Journal of Technology in Human Services*, 19, 5–18.
- National Telecommunications and Information Administration. (2011). *Digital Nation: Expanding Internet Usage, NTIA Research Preview February 2011*.

- U.S. Department of Commerce, Washington, DC. Retrieved September 29, 2011 from http://www.ntia.doc.gov/files/ntia/publications/ntia_internet_use_report_febru
- Nosek, B. A., & Banaji, M. R. (2001). The Go/No-go Association Task. *Social Cognition*, 19, 625–664.
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). E-research: Ethics, security, design, and control in psychological research on the Internet. *Journal of Social Issues*, 58, 161–176.
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2012). *Project Implicit*. Retrieved March 20, 2012, from <https://implicit.harvard.edu/implicit/>
- Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M. *et al.* (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, 18, 36–88.
- O’Neil, K. M., & Penrod, S. D. (2001). Methodological variables in web-based research that may affect results: Sample type, monetary incentives, and personal information. *Behavior Research Methods, Instruments, & Computers*, 33, 226–233.
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45, 867–872.
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5, 411–419.
- Park, C. L., Armeli, S., & Tennen, H. (2004). Appraisal-coping goodness of fit: A daily Internet study. *Personality and Social Psychology Bulletin*, 30, 558–569.
- Park, N., Peterson, C., & Seligman, M. E. P. (2006). Character strengths fifty-four nations and the fifty US states. *The Journal of Positive Psychology*, 1, 118–129.
- Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, 28, 450–461.
- Pew Internet and American Life Project (2011). *Internet Adoption, 1995–2011*.

Retrieved March 27, 2012, from <http://www.pewinternet.org/Trend-Data/Internet-Adoption.aspx>.

- Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, 49, 243–256.
- Reips, U.-D., & Neuhaus, C. (2002). WEXTOR: A web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, Instruments, & Computers*, 34, 234–240.
- Reis, H. T., Smith, S. M., Carmichael, C. L., Caprariello, P. A., Tsai, F. F., Rodrigues, A., & Maniaci, M. R. (2010). Are you happy for me? How sharing positive events with others provides personal and interpersonal benefits. *Journal of Personality and Social Psychology*, 99, 311–329.
- Rentfrow, P. J., Gosling, S. D., & Potter, J. (2008). A theory of the emergence, persistence, and expression of geographic variation in psychological characteristics. *Perspectives on Psychological Science*, 3, 339–369.
- Rosen, L. D., Cheever, N. A., Cummings, C., & Felt, J. (2008). The impact of emotionality and self-disclosure on online dating versus traditional dating. *Computers in Human Behavior*, 24, 2124–2157.
- Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, 58, 646–656.
- Saavedra, M. C., Chapman, K. E., & Rogge, R. D. (2010). Clarifying links between attachment and relationship quality: Hostile conflict and mindfulness as moderators. *Journal of Family Psychology*, 24, 380–390.
- Scherpenzeel, A. C., & Das, M. (2011). “True” longitudinal and probability-based Internet panels: Evidence from the Netherlands. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the Internet: Advances in applied methods and research strategies* (pp. 77–104). New York: Routledge.
- Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51, 515–530.
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309–316.

- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366.
- Simon, B., & Ruhs, D. (2008). Identity and politicization among Turkish migrants in Germany: The role of dual identification. *Journal of Personality and Social Psychology*, 95, 1354–1366.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.
- Skitka, L. J., & Sargis, E. G. (2006). The Internet as psychological laboratory. *Annual Review of Psychology*, 57, 529--555.
- Smyth, J. D., & Pearson, J. E. (2011). Internet survey methods: A review of strengths, weaknesses, and innovations. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the Internet: Advances in applied methods and research strategies* (pp. 11–44). New York: Routledge.
- Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of big five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology*, 94, 718–737.
- Stieger, S., Göritz, A. S., & Voracek, M. (2011). Handle with care: The impact of using Java applets in web-based studies on dropout and sample composition. *Cyberpsychology, Behavior, and Social Networking*, 14, 327–330.
- Stitka, L. J., & Sargis, E. G. (2006). The Internet as psychological laboratory. *Annual Review of Psychology*, 57, 529–555.
- Sullivan, D., Landau, M. J., Branscombe, N. R., & Rothschild, Z. K. (2012). Competitive victimhood as a response to accusations of ingroup harm doing. *Journal of Personality and Social Psychology*, 102, 778–795.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.
- Thiele, O., & Kaczmirek, L. (2010). Security and data protection: Collection, storage, and feedback in Internet research. In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (pp. 235–253). Washington, DC: American Psychological Association.

- Trepte, S., Reinecke, L., & Juechems, K. (2012). The social side of gaming: How playing online computer games creates online and offline social support. *Computers in Human Behavior*, 28, 832–839.
- Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research*, 23, 3–43.
- Williams, K. D., Cheung, C. K. T., & Choi, W. (2000). Cyberostracism: Effects of being ignored over the Internet. *Journal of Personality and Social Psychology*, 79, 748–762.
- Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28, 186–191.
- Wright, M. F., & Li, Y. (2011). The associations between young adults' face-to-face prosocial behavior and their online prosocial behaviors. *Computers in Human Behavior*, 27, 1959–1962.

¹ Internet recruitment can be distinguished from Internet data collection. In fact, Skitka and Sargis (2006) reported that 25% of the studies using Internet methods included in their review used the Internet to collect data from student samples. Our focus is on studies that make full use of the Internet for both recruitment and data collection.

² Internet methods have been combined with probability-based sampling strategies to create nationally representative samples (e.g., Scherpenzeel & Das, 2011). Conversely, the Internet is sometimes used to collect data from student samples (Skitka & Sargis, 2006). Therefore, the representativeness of samples depends more on specific recruitment strategies than on whether or not a study uses the Internet.

³ Although the technical details of writing HTML code, setting up servers, and implementing programming languages like Flash and JavaScript are beyond the scope of this chapter, numerous comprehensive resources are available for learning these skills. Several books and free online resources (e.g.,

<http://www.w3schools.com>) provide basic introductions to HTML, JavaScript, and other server-and client-side programming languages mentioned here. Some websites include sample programming that can be easily modified and added to study webpages (e.g., software to implement sliding response bars). There are also books more exclusively focused on the needs of online researchers (e.g., Fraley, 2004).

⁴ Data security concerns are particularly important for researchers hosting a study on their own server, as they must take precautions to ensure that the server (and the data it stores) is protected from attacks. Thiele and Kaczmarek (2010) provide a more detailed discussion of security precautions in Internet-based research.

⁵ Hardware differences can also affect the presentation of stimuli, leading to slightly more variability in presentation times for subliminal primes than would be possible in a carefully calibrated laboratory study. In addition, hardware and environmental factors (e.g., lighting, glare) can affect the appearance of images with regards to brightness, contrast, resolution, color, and size.

⁶ It should be noted that MTurk allows researchers to restrict participation to a specific country. Without enabling this option, it is likely that a large portion of respondents will come from India, which constitutes a sizable minority of MTurk workers.

Part three Data Analytic Strategies

Chapter eighteen Measurement

Reliability, Construct Validation, and Scale Construction Oliver P. John and Veronica Benet-Martínez*

Is an alpha reliability of .70 high enough? How do I know my questionnaire scale is unidimensional? What do I need to do to show that my measure is valid? These are the kinds of questions that methodologists are often asked. The answers to these questions are important for everybody who does empirical research in social-personality psychology, and they all involve basic issues in measurement. Yet when we teach courses on measurement and test construction, we seldom encounter much enthusiasm. In fact, most students think that measurement is outright boring. However, without measurement there would be no empirical science.

Consider an unusual and extreme but illustrative example: research on the vast societal problem of child molestation (see Harris & Rice, 1996). Theory suggested a crucial variable: Child molesters may sexually prefer and be responsive to children, whereas non-molesters are more responsive to adults. The researchers were faced with what seemed to be insurmountable measurement problems – how could they measure a construct such as “sexual responsiveness to children” in individuals who had reason to deny to others and themselves that they are attracted to children? Self-report did not seem a viable option when studying sex offenders and sexual aggression. Eventually, the researchers developed an ingenious phallometric procedure that allowed them to measure genital blood flow in response to slides depicting pictures of nude adults and children. Even though they had a seemingly fool-proof physiological measure, the investigators took painstaking care to attend to measurement issues: Do the blood flow measurements generalize across equivalent kinds of pictures? Do they replicate over time and testing situations? Do they validly differentiate groups of known offenders from (presumed) non-offenders? Do they converge with measures obtained with other methods, such as reports from clinicians, and would they predict future abuse?

Some General Considerations in Measurement

These questions – traditionally discussed under the headings of reliability and validity – all illustrate the fundamental concern of empirical science with *generalizability*, that is, the degree to which we can make inferences from our measurements or observations to other samples, items, measures, methods, outcomes, and so on (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). If we cannot make such generalizations, our measurements are obviously much less useful than if we can provide explicit evidence for generalizability.

The notion of generalizability also reminds us that good measurement implies not only that we can reproduce or replicate the same measurement but also that we can trust that the measurement has a particular meaning – we want to be able to make inferences about other variables that interest us. In the phallometric example, the blood flow measurements would be useless if they failed to help us understand differences between offenders and non-offenders. Another basic idea implicit in the idea of generalizability is that all psychological measurement – self-reports, observer ratings, even physiological measures – is prone to errors and that we cannot simply assume that a single measurement will generalize. Any one measurement may be distorted by numerous sources of error (e.g., there may have been something about the particular slide used, the blood flow meter may have shifted slightly, etc.), and the resulting observation (or score) is only imperfectly related to what we want to measure, namely sexual responsiveness. To counteract this limitation of single measurements, psychologists obtain multiple measurements (e.g., across different stimuli, experimenters, or observers) and then aggregate them into a more generalizable composite score.

Defining Measurement as Building and Evaluating Models

Admonished that psychologists often talk “at great length” about phenomena and concepts they have not defined (Dawes & Smith, 1985, p. 509), we shall briefly consider what measurement is and how it may be defined. An early definition comes from Stevens (1951), one of the founders of measurement theory, who suggested that measurement is the assignment of numbers to objects or events according to rules. However, it is now generally agreed that measurement requires more than that, and Dawes, Smith, and Himmelfarb (1993), and Judd and McClelland (1998) present excellent discussions of the relevant historical and conceptual issues. We agree that it is most useful to think of measurement as the process of building models that represent the phenomena of interest, typically in quantitative form. Judd and McClelland (1998, p. 181) articulated

this point of view:

The raw data...of the social and behavioral sciences...consist of infinitely minute observations of ongoing behavior and attributes of individuals, social groups, social environments, and other entities or objects that populate the social world. Measurement is the process by which these infinitely varied observations are reduced to compact descriptions or *models* that are presumed to represent meaningful regularities in the entities that are observed....

Accordingly, measurement consists of rules that assign scale or variable values to entities to represent the constructs that are thought to be theoretically meaningful. (emphasis added)

Like most models, measurement models (e.g., tests, scales, or variables) have to be reductions or simplifications to be useful. Although they should represent the best possible approximation of the phenomena of interest, we must expect them, like all “working models,” to be eventually proven wrong and to be superseded by successively better models. For this reason, measurement models must be specified explicitly so that they can be evaluated, disconfirmed, and improved. Moreover, we should not ask whether a particular model is true or correct; instead, we should build several plausible alternative models and ask: Given everything we know, which models can we rule out and which model is best at representing our data? Or, even more clearly: Which model is the least wrong? This kind of comparative model-testing (e.g., Judd, McClelland, & Culhane, 1995) is the best strategy for evaluating and improving our measurement procedures.

Psychometric and Representational Approaches to Measurement

The present chapter focuses on what has become known as the psychometric or nonrepresentational approach to measurement. Representational measurement has been discussed in several extensive reviews, especially in the context of attitude measurement (Dawes & Smith, 1985; Himmelfarb, 1993). In brief, the basic assumption of representational measurement is that numbers are assigned to entities such that the properties of the numbers (e.g., “greater than,” “multiplication”) represent empirical relations. A good example of representational measurement is the Mohs Scale of Hardness, which measures

the hardness of rocks in terms of an ordinal scale (Dawes & Smith, 1985): Rock X is harder than Rock Y if and only if X can scratch Y. The key feature of representational measurement in this example is the empirical relation that can be shown to exist between any pair of rocks and that can be represented by the “greater-than” relation among real numbers. One advantage of these kinds of measurement models is that they make predictions about the behavior of the individual entities being measured and thus provide internal consistency checks that can be used to disconfirm the model. For example, the ordinal hardness scale for rocks has to follow the transitivity rule, such that if Rock X is harder than Rock Y and Y is harder than Z, then X has to be harder than Z, and this prediction can be verified empirically by checking whether X does indeed scratch Z.

In contrast, the psychometric approach does not afford such internal consistency checks. For example, although the responses participants make on rating scales are often assigned numbers (e.g., 1 = *disagree strongly* and 5 = *agree strongly*), these numbers are not imbued with strong representational meaning that would permit consistency checks (see Dawes & Smith, 1985 for samples and a discussion of rating scales). Instead, the psychometric approach relies on aggregate patterns of data to evaluate a proposed measurement model. It does so because it assumes that each individual response or observation is so prone to error that consistency checks at this level of measurement are simply not meaningful and informative. For example, consider the two self-report items “I am a generous person” and “I am a stingy person” (Hampson, 1998). Although responses to these two items tend to be negatively correlated (i.e., most respondents claim one of the two traits but not both), the correlations are not even close to -1.0 , the number we would expect if people were semantically consistent. Instead, only some people are very consistent; there are vast individual differences, even among college students, and more verbally intelligent students show greater consistency (Goldberg & Kilkowski, 1984). In other words, people are not like rocks – they are much less consistent in their behavior, scratching or otherwise. Thus, the psychometric approach tends to ignore consistency checks at the level of the individual and instead relies on patterns of variances and covariances that reflect relations at the aggregate level in probabilistic form (e.g., in this sample, individuals who gave relatively high ratings to “generous” were unlikely to give high ratings to “stingy”).

Although representational measurement promised to provide a strong and defensible foundation for psychological measurement, it has so far failed to deliver on that promise. During the 1970s and 1980s a slew of studies, inspired

by Tversky and Kahneman's (1974) pioneering work, showed that people's preferences, risk perceptions, political attitudes, and so on often violate the transitivity rule required for ordinal scaling and that judgments may shift substantially depending on the framing of the questions or items. Dawes and Smith (1985) noted that “representational measurement is rare in the field of attitude; instead, this field is permeated by questionnaires and rating scales” (pp. 511–512). Somewhat more recently, Cliff (1992) called representational measurement “the revolution that never happened” (p. 186), and Dawes (1994) concurred. For these reasons, and because Judd and McClelland (1998) provided an excellent up-to-date review and discussion of the representational approach, the present chapter is devoted to the psychometric approach.

Overview

The remainder of this chapter is organized into three parts. We begin with the historically early conceptions of reliability and then move to more recent and increasingly complex views that emphasize construct validation and model-testing as a broader, more integrative approach. In the first part of this chapter we consider issues traditionally discussed under the heading of reliability, review several still persistent definitions or “types” of reliability coefficients, discuss in some detail the problems and misuses of coefficient alpha, the most commonly used psychometric index in social-personality psychology, and then suggest generalizability theory as a broader and more heuristic perspective. In the second part we examine issues related to construct validation, beginning with early definitions and designs to establish validity, followed by a broader view that considers construct validation as the crucial issue in psychological measurement and includes a broad range of validity evidence, focusing on convergent and discriminant aspects. In the third part we consider model-testing in construct validation and scale construction. After a brief introduction to measurement models in structural equation modeling (SEM), we discuss an empirical example that reexamines the issue of dimensionality as an aspect of structural validity, and then we consider issues in questionnaire construction, reviewing three classical strategies (external-criterion, rational-intuitive, and internal-factor analytic) and suggesting an integrated model adopting the construct-oriented approach.

It sometimes seems that methodologists write papers that are of great interest to other methodologists. Instead, the present chapter is focused on what in our experience has proven useful and of interest to graduate and postdoctoral

students, with the goal of devising sound measurement models and evaluating them, rather than covering mathematical formulae or statistical derivations. We discuss current practice, even when it is outmoded, and then point to more recent conceptualizations. Finally, whenever possible, we have avoided technical language, omitted Greek symbols, and used examples to make this chapter as concrete and accessible as possible.

Reliability and Generalizability

It should by now be obvious that most measurement procedures in psychology are subject to “error.” Many different sources may contribute to such error. In the social-personality literature, the observations, ratings, or judgments that constitute the measurement procedure are typically made by humans who are subject to a wide range of frailties. Research participants may become careless or inattentive, bored or fatigued, and may not always be motivated to do their best. The particular conditions and points in time when ratings are made or recorded may also contribute error. Further errors may be introduced by the rating or recording forms given to the raters to obtain their judgments; the instructions, definitions, and questions on these forms may be difficult to understand or require complex discriminations, again entering error into the measurement. In short, characteristics of the participant, the testing situation, the test or instrument, and the experimenter can all affect reliability.

Reliability refers to the consistency of a measurement procedure, and indices of reliability describe the extent to which the scores produced by the measurement procedure are reproducible. Consider the example of a bathroom scale; if it gives different readings in three successive weighings of the same person, we would hardly call the scale reliable.

Classical Test Theory

Issues of reliability have traditionally been treated within the framework of classical test theory (Gulliksen, 1950; Lord & Novick, 1968). If a given measurement X is subject to error e , then the measurement without the error, $X - e$, would represent the accurate or “true” measurement T (e.g., the person's actual weight). This seemingly simple formulation, that each measurement can be partitioned into a true score, T , and measurement error, e , is the fundamental assumption of classic test theory. Conceptually, each true score represents the mean of a very large number of measurements on a specific individual, whereas

measurement error represents all of the momentary variations in the circumstances of measurement that are unrelated to the measurement procedure itself. Such errors are assumed to be random (a rather strong assumption to which we will return), and it is this assumption that permits the definition of error in statistical terms.

All conceptions of reliability involve the notion of repeated measurements. Classical test theory has relied heavily on the notion of *parallel tests* – that is, two tests that have the same mean, variance, and distributional characteristics, and that correlate equally with external variables (Lord & Novick, 1968). Under these assumptions, true score and measurement error can be treated as independent. It follows that the variance of the observed scores equals the sum of the variance of the true scores and the variance of the measurement error:

$$\text{Variance } (X) = \text{Variance } (T + e)$$

$$= \text{Variance } (T) + \text{Variance } (e).$$

Reliability can then be defined as the ratio of the true-score variance to the observed-score variance, which is equivalent to 1 minus the ratio of error variance to observed-score variance:

$$\text{Reliability} = \text{Variance } (T) / \text{Variance } (X)$$

$$= 1 - [\text{Variance } (e) / \text{Variance } (X)].$$

In other words, if there is no error, reliability would be 1; if there is only error and no true-score variance, reliability would be 0. The correlation between the observed variable and the true score is the square root of the reliability.

Specific Types of Reliability Evidence

Because classical test theory defined parallel tests in purely mathematical terms, it provided little substantive specification or restriction of the types of measurement procedures that might be considered parallel. Beginning in the 1950s, several designs were distinguished, and they are summarized in [Table 18.1](#): retest (or stability), equivalence, and internal consistency (or split-half). These distinctions were meant to convey the idea that “reliability is a generic term referring to many types of evidence” (American Psychological Association, 1954, p. 28). Clearly, the different designs spelled out in [Table 18.1](#) take into account quite different sources of error. *Retest* (or stability) designs estimate how much responses vary within individuals across time and situation, thus reflecting error resulting from differences in the situation and conditions of test administration or observation.¹ *Equivalence* procedures estimate error

attributable to different content sampling and item selection in two alternate forms of the test. *Internal-consistency* procedures offer an estimate of error associated with the particular selection of items; error is high (and internal consistency is low) when items are heterogeneous in content and lack content saturation.

Table 18.1. *Reliability: Facets of Generalizability, Traditional Definitions of Reliability Coefficients, and Estimation Procedures*

Facet of Generalizability	Major Sources of Error	Traditional Reliability Coefficient	Procedure	Statistical Analysis
Times	Change of participant's responses over time; change in testing situation	Retest (or stability)	Test participants at different times with same form	Pearson or intraclass correlation
Forms	Differences in content sampling across "parallel" forms	Equivalence	Test participants at one time with two forms covering same content	Pearson or intraclass correlation
Items	Content heterogeneity and low content saturation in the items	(a) Split-half (b) Internal consistency	Test participants with multiple items at one time	(a) Correlation between test halves (Spearman-Brown corrected) (b) Coefficient alpha
Judges or observers	Disagreement among judges	Internal consistency	Obtain ratings from multiple judges on one form and occasion	(a) Pairwise interjudge correlation (b) Coefficient alpha (c) Intraclass correlation

Coefficient Alpha: Ubiquitous but Not a Panacea

We now consider coefficient alpha (Cronbach, 1951) because this internal-consistency index plays such an important role in the social-personality literature. A perusal of the articles published in the leading social-personality journals shows that de facto alpha is the index of choice when authors want to claim that their measure is reliable. Moreover, contrary even to the recommendations in the *Standards* (American Psychological Association, 1985), alpha is usually the only reliability evidence considered.

Why has alpha become the golden standard of measurement reliability? We suspect it is the relative ease with which alpha is both obtained and computed. Alpha does not require collecting data at two different times from the same participants, as retest reliability would, or the construction of two alternate forms of a scale, as parallel-form reliability would require. Alpha is the "least effort"

reliability index; it can be used as long as the same participants responded to multiple items thought to indicate the same construct. And computationally, today's statistical software packages allow the user to view the alpha of many alternative scales formed from any motley collection of items with just a few mouse clicks. However, although alpha has many important uses, it also has some important limitations. Although long known to methodologists, these limitations are often underappreciated by researchers and are therefore reviewed here in some detail.

Two Determinants of Alpha. Cronbach's (1951) alpha is a generalization of split-half reliability, representing the mean of the reliabilities computed from all possible split halves of the test. As such, alpha is a function of two parameters: (1) the interrelatedness of the items in a test or scale and (2) the length of the test. Consider Table 18.2, which shows the interitem correlation matrices for two hypothetical tests, one with 10 items and one with 6 items, constructed by Schmitt (1996).² Both tests have the same alpha of .81 but they achieve that alpha in two rather different ways. Test B has only 6 items but, on average, these items are more highly intercorrelated (mean $r = .42$) than the 10 items of Test A (mean $r = .33$).

Table 18.2. Interitem Correlation Matrices for Two Hypothetical Tests with the Same Coefficient Alpha Reliability of .81

Test A with 10 items											Test B with 6 items						
Variable	1	2	3	4	5	6	7	8	9	10	Variable	1	2	3	4	5	6
1	—										1	—					
2	.3	—									2	.6	—				
3	.3	.3	—								3	.6	.6	—			
4	.3	.3	.3	—							4	.3	.3	.3	—		
5	.3	.3	.3	.3	—						5	.3	.3	.3	.6	—	
6	.3	.3	.3	.3	.3	—					6	.3	.3	.3	.6	.6	—
7	.3	.3	.3	.3	.3	.3	—										
8	.3	.3	.3	.3	.3	.3	.3	—									
9	.3	.3	.3	.3	.3	.3	.3	.3	—								
10	.3	.3	.3	.3	.3	.3	.3	.3	.3	—							

The idea that test length can compensate for lower levels of interitem correlation is formalized in the Spearman-Brown prophecy formula, which specifies the relation between test length and reliability (see, e.g., Lord & Novick, 1968). Given a particular level of mean interitem correlation, the Spearman-Brown formula allows the researcher to derive the number of items

needed to achieve a certain level of alpha. Figure 18.1 shows this relation for mean interitem correlations of .20, .40, .60, and .80 in graphic form. Three points are worth noting. First, the alpha reliability of the total scale always increases as the number of items increases (as long as adding items does not lower the mean interitem correlation). Second, the utility of adding ever more items diminishes quickly, so that adding the 15th item leads to a much lesser increase in alpha than adding the 5th item, just like consuming the 15th chocolate bar or beer adds less enjoyment than did the earlier ones. Third, less is to be gained from adding more items if those items are highly intercorrelated (e.g., .60) than when they show little content saturation (e.g., .20). The lesson here is that we need to be careful in interpreting alpha: We cannot interpret empirical findings without considering scale length. In contexts where the researcher is interested in the homogeneity of the items, a direct index of item content saturation (e.g., the mean interitem correlation) may be more informative than alpha.

Alpha Does Not Index Unidimensionality. The examples in Table 18.2 also illustrate a second issue with alpha: Contrary to popular belief, alpha does not measure the homogeneity of the interitem intercorrelations, nor does it indicate that a scale is unidimensional. In fact, although Tests A and B in Table 18.2 have the same alpha, they differ radically in the homogeneity (vs. dispersion) of the correlations among their items. For Test A, they are completely homogeneous (all are .3, with a standard deviation (*SD*) of 0 in this hypothetical example), whereas for Test B they vary considerably (from .3 to .6, with an *SD* of .15). Because alpha does not represent this variability, Cortina (1993) derived an index that reflects the spread of interitem correlations and argued that this index should be reported along with alpha. A large spread in interitem correlations is a bad sign because it suggests that either the test is multidimensional or the interitem correlations are distorted by substantial sampling error.³ In the example, the pattern of item intercorrelations for Test B suggests that the problem here is multidimensionality. Clearly, the responses to these six items are a function not of one, but two, factors: Items 1, 2, and 3 correlate much more substantially (mean $r = .6$) with each other than they correlate (mean $r = .3$) with items 4, 5, and 6, which in turn correlate more highly with each other (mean $r = .6$). Alpha disguises this rather important difference between Tests A and B.

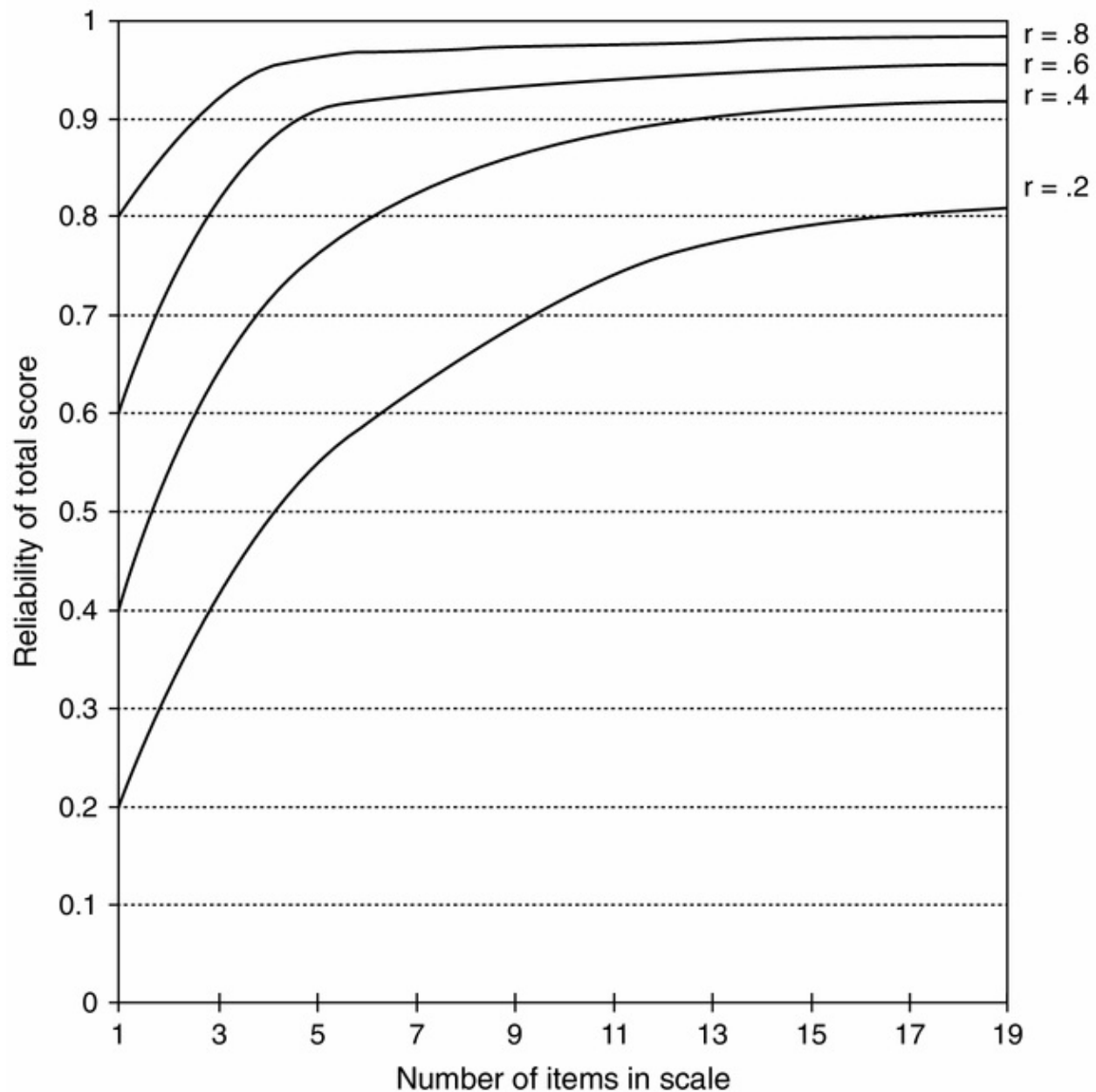


Figure 18.1. Alpha reliability as a function of the number of items included in the scale at four levels of mean interitem correlation.

Because alpha cannot address it, unidimensionality needs to be established in other ways. The most rigorous approach is to use the confirmatory factor analysis part of SEM (Jöreskog & Sörbom, 1981; see also Bentler, 1980; Fabrigar & Wegener, Chapter 19 in this volume; Widaman & Grimm, Chapter 20 in this volume). SEM allows us to test how well the interitem correlation matrix fits a single-factor, rather than multifactor, model. In other words, how well can the loadings on a single factor reproduce the correlation matrix actually

observed? Not surprisingly in these artificial data, the SEM results show that the one-factor model describes the data pattern for Test A perfectly; all items have a factor loading of .548 (i.e., the square root of .3, which is the size of all interitem correlations in this example) and an error term of .837. In contrast, for Test B the fit of the one-factor model was unacceptable even though the item loadings of .648 (i.e., the square root of .42, which is the mean interitem correlation) were higher than for the truly unidimensional Test A. As expected, a two-factor model significantly increased fit for Test B, and perfect fit was obtained when we specified a model with two correlated factors. Reflecting their .60 correlation with each other, items 1, 2, and 3 loaded .775 on factor 1 and 0 on factor 2, whereas items 4, 5, and 6 loaded 0 on factor 1 and .775 on factor 2; the mean intercorrelation of .3 between the items in these two sets gave rise to a .50 correlation estimated between the two latent factors. It is important to emphasize that the issue of error (or unreliability) present in an item is separate from the issue of multidimensionality (which is discussed in later sections on structural validity). In the SEM analyses summarized earlier, the item loadings represent how much of the item variance is shared across items (thus generalizable), whereas error is captured by the residual item variance (i.e., 1 minus the squared loading), indicating how much variance is unique to that item; the proportion of shared to total item variance is often referred to as *content saturation*. Dimensionality, on the other hand, is captured by the relative fit of the one-factor model over multiple-factor models. Thus, comparing again Tests A and B in [Table 18.2](#), the longer Test A is clearly more unidimensional than Test B is, yet its items do not show greater content saturation (i.e., higher factor loadings and lower error terms). In other words, unidimensionality does not imply lower levels of measurement error (i.e., unreliability) and vice versa. We return to these issues later in this chapter when we discuss the measurement model in SEM.

Once we know that a test is multidimensional, can we go ahead and still use alpha as a reliability index? Unfortunately, the answer is no. As Cronbach (1947, 1951) recognized early on, we can estimate the reliability of a multidimensional test or scale only through parallel forms, and the two parallel forms must show the same factor structure. In fact, if the test is not unidimensional, alpha underestimates reliability (see Schmitt, 1996 for an example). Thus, if a test is found to be multidimensional, one should score two unidimensional subscales and then use alpha to index their reliabilities separately.⁴

How Large Should Alpha Be? It Depends on the Construct. Students often ask questions like “my scale has an alpha of .70 – isn't that good enough?” and

they are frustrated when the answer is “that depends.” Although it would be nice to have a simple cookbook for measurement decisions, there is no particular level of alpha that is necessary, adequate, or even desirable in all contexts. Although Nunnally (1978) suggested that “reliabilities of .7 or higher will suffice” (p. 245), an alpha of .70 is not a benchmark every scale must pass. It is easy to find examples in the literature that use this arbitrary standard. For example, Gray-Little, Williams, and Hancock (1997) noted that alphas between .72 and .88 are usually taken to indicate “acceptable to high reliability” (p. 444). However, as we have seen above, alpha needs to be interpreted in terms of its two main parameters – interitem correlation and scale length – and in the context of how these two parameters fit the nature and definition of the construct to be measured. In any one context, a particular alpha may be just right, or too low, or too high. As Pedhazur and Schmelkin (1991) put it,

Does a .5 reliability coefficient stink? To answer this question, no authoritative source will do. Rather, *it is for the user to determine what amount of error variance he or she is willing to tolerate, given the specific circumstances of the study.*

(p. 110)

The definition of the construct to be measured is a crucial parameter in interpreting alpha. Consider a researcher who wants to measure the broad construct of extraversion, which includes sociability, assertiveness, and talkativeness, and has constructed a scale with the following items: “I like to go to parties,” “Parties are a lot of fun for me,” “I do not enjoy parties,” (reverse scored), and “I’d rather go to a party than spend the evening alone.” Note that these items are essentially paraphrases of each other and represent the same item content (liking parties) stated in slightly different ways. Cattell (1972) called these kinds of scales “bloated specifics” – they have high alphas simply because the item content is so redundant and interitem correlations are very high. Thus, alphas in the high .80s or even .90s, especially for short scales, may not indicate an impressively reliable scale but instead signal redundancy or narrowness in item content.

For example, the 10-item Rosenberg (1979) Self-Esteem Scale has alphas approaching .90 in student samples, and the pairwise correlations between some items approach .70 (Gray-Little et al., 1997). Some of these self-esteem items turn out to be almost synonymous, such as “I certainly feel useless at times” and “At times I think I am no good at all.” Although such redundant items increase

alpha, they do not add unique (and thus incremental) information and can often be omitted in the interest of efficiency, suggesting that the scale can be abbreviated without much loss of information (see Robins & Hendin, 1999).

This phenomenon is also known as the *attenuation paradox* because increasing the internal consistency of a test beyond a certain point will not enhance construct validity and may even come at the expense of validity when the added items emphasize one part of the construct (e.g., party-going) over other important parts (e.g., assertiveness). This paradox emphasizes an important point in this chapter: Our goal in measurement is to maximize validity rather than internal consistency, and issues of meaning and conceptualization play a key role in all decisions about measurement.

The party-going items on our imaginary extraversion scale illustrate how easy it is to boost alpha by adding redundant items to a scale. However, unless one is specifically interested in party-going behavior, this strategy is not very useful: The narrow content representation (i.e., high content homogeneity) would make this scale less useful as a measure of sociability and even less useful as a measure of the broader construct of extraversion. Although the scale may predict the frequency of party attendance with great precision (or fidelity), it is less likely to relate to anything else of interest because of its narrow bandwidth. Conversely, broad-bandwidth measures (e.g., an extraversion scale or a general attitude measure of conservatism) can predict a wider range of outcomes or behaviors but do so with lower fidelity. This phenomenon is known as the bandwidth-fidelity trade-off (Cronbach & Gleser, 1957) and has proven to be of considerable importance in the literature on attitudes (Eagly & Chaiken, 1993; Fishbein & Ajzen, 1974) and personality traits (Epstein, 1980; John, Hampson, & Goldberg, 1991). In general, then, the attitude or trait serving as the predictor should be measured at a similar level of abstraction as the criterion to be predicted, so that predictive relations are not going to be attenuated.

The close connection between the hierarchical level of the construct to be measured and the content homogeneity of the items is illustrated in Figure 18.2. Sociability, assertiveness, and talkativeness are three scales that are positively intercorrelated and together define the broader construct of extraversion. (Our earlier example of the party-going scale might be represented as an even lower-level scale, representing one of the components of the sociability scale.) Consider the assertiveness scale. Because its six items are selected to represent a narrow range of content (e.g., all the assertiveness items have to do with dominance and self-assertion in various situations), item content should be

relatively homogeneous, leading to a substantial mean interitem correlation. Now consider an equally long extraversion scale (shown on the right side of Figure 18.2), made up of two sociability items, two assertiveness items, and two talkativeness items. Compared with the lower-level scales, the item content on this scale is much more heterogeneous, leading to a lower mean interitem correlation and thus a lower alpha.

One implication of Figure 18.2 is that if one wants to measure broader constructs, one should probably include a larger number of items to compensate for the greater content heterogeneity; for example, one might use all 18 items to measure the superordinate extraversion construct defined on the left side of Figure 18.2.

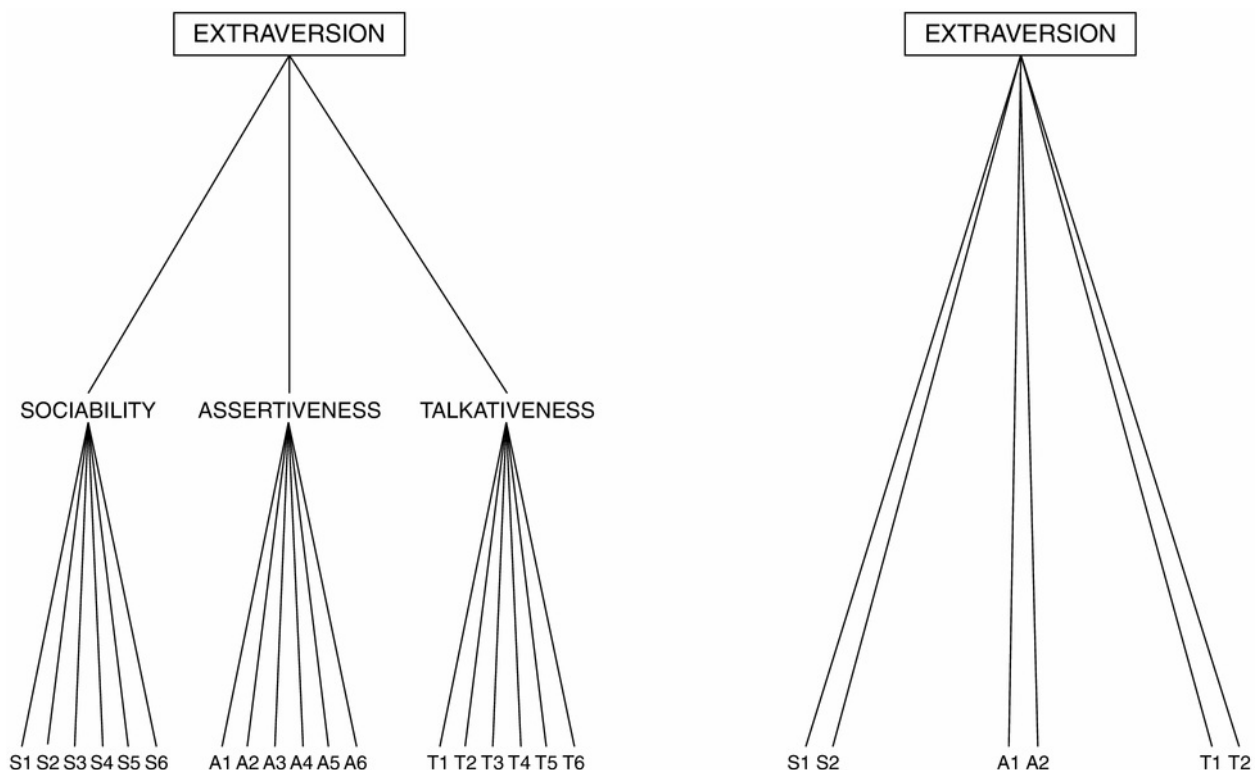


Figure 18.2. Relation between hierarchical level and content homogeneity in interpreting the size of alpha reliability.

This principle underlies the construction of many hierarchically organized assessment instruments, such as in the literature on the self-concept (e.g., Marsh, Byrne, & Shavelson, 1992) and on personality traits. For example, in Costa and McCrae's (1992) NEO PI-R, each of the Big Five personality dimensions is defined by six more specific “facet” scales, which, in turn, are each measured with eight items; the resulting 48-item superordinate Big Five scales all have

reliabilities exceeding .90.

Correcting for Attenuation

According to classical test theory (e.g., Lord & Novick, 1968), researchers should be concerned about reliability because the reliability of a measure constrains how strongly that measure may correlate with another variable (e.g., an external criterion). If error is truly random, as classical test theory assumes, the upper limit of the correlation for a measure is not 1.0 but the square root of its reliability (i.e., the correlation of the measure with itself). Thus, the true correlation between the measure and another variable may be underestimated (i.e., attenuated) when reliability is inadequate. For example, for a reliability of .70, the expected upper limit for a correlation would be the square root of .70, namely .84, and for a reliability as low as .60, the limit would be .77. These numerical examples show that lower reliabilities can reduce estimates of external correlations and, everything else being equal, higher reliabilities are to be preferred.

However, other properties of the instrument need to be considered as well in planning one's research. As Burisch (1984, 1986) has shown, “short scales not only save testing time but also avoid subject boredom and fatigue....There are subjects...from whom you won't get any response if the test looks too long” (p. 112). Consider, for example, the 48-item long Big Five trait scales from the NEO PI-R (Costa & McCrae, 1992). Despite their reliabilities above .90, they are used less frequently by other researchers than are shorter scales that have reliabilities in the .70s and .80s. When participant time and attention is at a premium, the trade-off between length and reliability may well be worth it – note that the drop in reliabilities from .90 to .80 and to .70 lowers the upper limits for external correlations only from .95 to .89 and to .84 for the shorter scales.

Researchers sometimes use reliability indices (typically alpha) to correct observed correlations between two measures for attenuation attributable to unreliability. Such corrections are sometimes used to estimate the correlation between the latent constructs underlying their measures (see also the section on SEM later in the chapter) – that is, what would the correlation be if both measures were assessed with perfect reliability? This can be useful to compare effect sizes across variables or studies. Another application is in contexts where researchers want to distinguish the long-term stability of attitudes or personality from the reliability of measurement or compare stability estimates for different groups, such as men and women (Block & Robins, 1993). The correction

formula (Cohen & Cohen, 1975; Lord & Novick, 1968) is simple: divide the observed correlation by the square root of the product of the two reliabilities. This correction expresses the size of the association relative to the maximum correlation attainable given the imperfect reliabilities of the two measures.

However, the ease of this correction should not lead to sloppy measurement. Appealing to the relative brevity of one's measures to excuse low reliability is, as we have seen from the earlier discussion, not the only explanation for low alphas, and certainly not an excuse. In many situations, low reliability will create problems for estimating effect sizes and testing hypotheses. This is especially true in multivariate applications, such as multitrait multimethod matrices (discussed later), where unequal reliabilities might bias conclusions about convergent and discriminant validity (West & Finch, 1997). In general, then, researchers are well advised to invest the time and effort needed to construct reliable measures.

Reporting Basic Psychometric Data

Whereas most social-personality researchers do report alpha reliabilities for their scales, few report their intercorrelations. This information is often crucial, for example, when multiple scales are scored from the same data source (e.g., a self-report attitude measure) and when the research question implies relative independence among the constructs measured. For example, intercorrelations among predictors are important for understanding the results of multiple regression analyses (e.g., see the numerical examples provided by Goldberg, 1991), and concerns have been raised about the correlations among constructs postulated to be conceptually unrelated, such as the Big Five personality dimensions (e.g., Block, 1995; John & Srivastava, 1999). Thus, we agree with Schmitt (1996), who argued that, at the very least, research reports should regularly present a matrix that includes reliability information for the key measures (on the diagonal), the intercorrelations among these measures (below the diagonal), and probably also the intercorrelations corrected for attenuation resulting from unreliability (above the diagonal). Table 18.3 gives an example from our own research on the Big Five Inventory scales (John, Donahue, & Kentle, 1991; John & Srivastava, 1999). The inclusion of the uncorrected intercorrelations allows the reader to evaluate the size of the reliability coefficients relative to the overlap among the scales; reliabilities should be substantially larger than these intercorrelations. The inclusion of corrected intercorrelations is helpful because it removes differences among the

intercorrelations that are attributable simply to differential reliability, thus making comparisons among intercorrelations much easier. The corrected coefficients are also useful for identifying pairs of scales that lack discriminance – that is, they are so highly intercorrelated that postulating two separate underlying constructs is not sensible either theoretically or practically.

Table 18.3. *How to Report Simple Psychometric Information for Multiple Scales: Alpha Coefficients, Observed Correlations, and Corrected Correlations Among the Big Five Inventory (BFI) Scales*

Scales	E	A	C	N	O
Extraversion (E)	(.88)	.17	.28	– .34	.30
Agreeableness (A)	.14	(.79)	.34	– .38	.06
Conscientiousness (C)	.24	.27	(.82)	– .22	.10
Neuroticism (N)	– .29	– .31	– .18	(.84)	– .17
Openness (O)	.25	.05	.08	– .14	(.81)

Note: N = 711 U.S. college students. Data from Benet-Martínez and John (1998). Alpha coefficients are presented on the diagonal, observed correlations below the diagonal, and correlations corrected for attenuation above the diagonal.

Beyond Classical Test Theory: Generalizability Theory

The distinctions among “types of reliability” emphasized in the literature and summarized in [Table 18.1](#) had a number of unfortunate consequences. First, they masked a major shortcoming of classical test theory: If all these measures were indeed parallel and all errors truly random, then all these approaches to reliability should yield the same answer. Unfortunately, they do not; reliability depends on the particular facet of generalization being examined (Cronbach, Rajaratnam, & Gleser, 1963). Second, what had been intended as heuristic distinctions became reified as “the Stability Coefficient” or “the Alpha Coefficient,” even though the notion of reliability was intended as a general concept. Third, the classification itself was too simple, equating particular kinds of reliability evidence with only one source of error and resulting in a restrictive terminology that cannot fully capture the broad range and combination of

multiple error sources that are of interest in most research and measurement applications (Shavelson, Webb, & Rowley, 1989).

Therefore, the American Psychological Association (APA) (1985) recommended in subsequent editions of the *Standards for Educational and Psychological Testing* that these distinctions and terminology be abolished and replaced by the broader view advocated by generalizability theory (Cronbach et al., 1963). Regrettably, however, practice has not changed sufficiently over the years, and generalizability theory has not fully replaced these more simplistic notions. To emphasize that the classical conception of random error is outdated, the very first column in our Table 18.1 spells out the facet of generalizability that is being varied and studied in each of these generalizability designs.

Generalizability theory holds that we are interested in the “reliability” of an observation or measurement because we wish to generalize from this observation to some other class of observations. For example, as shown by the last row in Table 18.1, concern with interjudge agreement may actually be a concern with the question of how accurately we can generalize from a given set of ratings to ratings by another set of judges. Or we might want to know how well scores on an attitude scale constructed according to one set of procedures generalize to another scale constructed according to different procedures. Or we might want to test the generalizability of a scale originally developed in English to a Chinese language and cultural context.

All these facets of generalizability represent legitimate research concerns that we will reconsider later in this chapter under the heading of construct validation; they can be studied systematically in generalizability designs, both individually and together. These designs allow the researcher to deliberately vary the facets that potentially influence observed scores and estimate the variance attributable to each facet (Cronbach et al., 1972). In other words, whereas classical test theory tries to estimate the portion of variance that is attributable to “error,” generalizability theory aims to estimate the extent to which specific sources of variance contribute to test scores under carefully defined conditions. Thus, instead of the traditional reliability coefficients listed in Table 18.1, we should use more general estimates, such as intraclass correlation coefficients (see McGraw & Wong, 1996; Shrout & Fleiss, 1979), to probe particular aspects of the dependability of measures. For example, the intraclass correlation coefficient (see Judd & McClelland, 1998 for numerical examples) can be used to index the generalizability of one set of judges to a universe of similar judges.

It is perplexing: Generalizability theory should hold considerable appeal for

social-personality psychologists because the extent to which we can generalize across items, instruments, contexts, groups, languages, and cultures is crucial to the claims we can make about our findings. Despite excellent and readable introductions (e.g., Shavelson et al., 1989), generalizability theory is not used as widely as it should be. A recent exception is the flourishing research on determinants of consensus among personality raters (Kenny, 1994; see also Kenny & Kashy, Chapter 22 in this volume) and the determinants of self-other agreement (John & Robins, 1993).

Generalizability theory is especially useful when data are collected in nested designs and multiple facets may influence reliability. A nice illustration is King and Figueredo's (1997) study of chimpanzee personality differences. They collected ratings of chimpanzees differing in age and sex (subject variables) on 40 traits (stimulus variables) at several different zoos (setting variables) from animal keepers familiar with the animals to varying degrees (observer variables). They then used a generalizability design to show how these facets affected agreement among the judges; fortunately for their purposes, setting and subject variables turned out to be unimportant.

Item Response Theory

The measurement model and procedures of classical test theory have also been criticized by psychometricians advocating item response theory (IRT) as an alternative and more advanced approach (Embretson, 1996; Mellenbergh, 1996). In the classical conception of reliability, the characteristics of the individual test-taker and the characteristics of the test cannot be separated (Hambleton, Swaminathan, & Rogers, 1991). That is, the person's standing (or level) on the underlying construct is defined only in terms of responses on the particular test; thus, the same person may appear quite liberal on a test that includes many items measuring extremely conservative beliefs but quite conservative on a test that includes many items measuring radical liberal beliefs. Furthermore, the psychometric characteristics of the test depend on the particular sample of respondents being measured; for example, whether a belief item from a conservatism scale reliably discriminates high and low scorers depends on the level of conservatism of the sample, so that the same test may work well in an undergraduate student sample but fail to make reliable distinctions among bible-belt evangelists. In short, classical test theory is not helpful if we want to compare individuals who have taken different tests measuring the same construct, or if we want to compare items answered by different groups of

individuals.

Another limitation of classical test theory is the assumption that the error of measurement is the same for all individuals in the sample – an implausible assumption given that tests and items differ in their ability to discriminate among respondents at different levels of the underlying construct (Lord, 1984; see Widaman & Grimm, Chapter 20 in this volume). Moreover, classical theory is test-oriented rather than item-oriented and thus does not make predictions about how an individual or group will perform on a particular item.

These limitations can be addressed in IRT, which describes the relation between individuals' responses to a particular item and the construct underlying those responses with a function called the *item characteristic curve*. This curve depicts the probability that individuals at different levels of the construct would endorse the item; it thus provides information about how well the item discriminates those with high versus low levels of the underlying trait and also about how difficult the item is. This information is particularly useful to researchers interested in detecting biases in their items; according to IRT, an item is an unbiased measure of a construct, say conservatism, if individuals who are equally conservative have the same expected score on the item, regardless of conceptually unrelated memberships in gender, ethnic, or cultural groups.

In the context of constructing and evaluating scales and other multi-item measures, IRT procedures have two attractive features. First, they permit researchers to select items on the basis of both difficulty and discrimination rather than relying on the item-total correlations offered by classical test theory. Second, IRT procedures can be used to assess a person's standing on the construct without having to administer the entire scale, a procedure known as computerized adaptive testing (Waller & Reise, 1989).

Until recently, IRT was limited computationally to dichotomous (true-false) response formats and unidimensional constructs. It was therefore much more useful for educational and achievement research (where item difficulty has an inherent psychological meaning) than for social-personality research, which relies heavily on multistep rating scales. However, as extensions of IRT and IRT software to rating scales and multidimensional models (Kelderman & Rijkes, 1994) become more accessible, IRT's "new rules of measurement" (Embretson, 1996, p. 341) are likely to appear more frequently in our journals. For example, Gray-Little *et al.* (1997) used IRT to explore the properties of the 10 items on the Rosenberg Self-Esteem Scale. Results indicated that the 10 items indeed define a unidimensional trait. However, given the uniformity of the item

discrimination parameters, the scale could easily be shortened without compromising the measurement of global self-esteem, a conclusion consistent with our earlier discussion of construct definitions and item redundancy (Robins & Hendin, 1999). The IRT analyses also indicated that the items discriminate better at low and moderate levels of self-esteem than they do at higher levels.

With its current selection of items, the scale may fail to differentiate reliably between truly high levels of self-esteem and narcissistically exaggerated, grandiose self-views (John & Robins, 1994).

More generally, then, IRT provides quantitative procedures to describe the relation of a particular item to the latent construct being measured in terms of difficulty and discrimination parameters. This information can be useful for item analysis and scale construction, permitting researchers to select items that best measure a particular level of the construct of interest and detecting items that are biased for particular groups of individuals.

To summarize, in this section we focused on classical test theory approaches to reliability, specific types of reliability indices, issues with coefficient alpha (test length, unidimensionality, and construct definitions), and the practice of correcting for attenuation. In discussing these issues we mentioned such concepts as latent (or underlying) constructs, construct definitions, dimensionality, criterion variables, and discriminant relations, but did not discuss them systematically. These concepts raise complex conceptional issues and highlight that the meaning and interpretation of measurements is crucial to evaluating the quality of our measurements. Traditionally, issues of score meaning and interpretation are discussed under the heading of validity. We focus on the validity of measured variables here; the validity of manipulated variables is discussed in [Chapters 2 and 3](#) of this volume.

Construct Validation

Traditional Definitions of Validity

As described by Cronbach and Meehl (1955), the APA committee on psychological tests initially distinguished among several types of validity, which are given in the top part of [Table 18.4](#). Content validity is established by demonstrating that the items are a representative sample of the universe of item content relevant to the construct. This aspect of validity is typically established deductively; first the investigator defines a universe of items (i.e., a hypothetical

set of all possible kinds of relevant item content) and then samples systematically from that universe to assemble the test items. Face validity concerns theoretical considerations about the appropriateness of the items, particularly whether they appear to assess attributes and behaviors relevant to the intended construct; that is, do the items look reasonable and sensible as indicators of the construct they are supposed to measure?

Table 18.4. *Types of Validity and Validity Evidence: Major Approaches*

Early Approaches (e.g., Cronbach & Meehl, [1955](#))

Content validity: Extent to which the items are a representative sample of the behavior domain to be measured

Face validity: Extent to which the items appear to measure the intended construct

Criterion-oriented (or external) validity:

(a) Predictive: Extent to which an individual's future score on a criterion is predicted from prior test scores

(b) Concurrent: Extent to which the test scores estimate an individual's present criterion score

Construct validity: Whether the measure accurately reflects the construct intended to measure

Elaboration of Construct Validity (e.g., Loevinger, [1957](#); Messick, [1989](#))

Content validity: Evidence of content relevance, representativeness, and technical quality of items

Substantive validity: Evidence for response consistencies or performance regularities that are reflective of domain processes

Structural validity: Evidence for internal structure of the scores that is

consistent with the internal structure of the construct domain

Generalizability: Evidence for score properties and interpretations that generalize to and across population groups, settings, and tasks

Consequential validity: Rationale and evidence for evaluating the intended and unintended consequences of score interpretation and use, including test bias and fairness

External validity: Convergent and discriminant evidence from multitrait multimethod comparisons, as well as criterion relevance

Examples of Validation Procedures

Expert judgments and review: Test whether experts agree that items are relevant and represent construct domain; use ratings to assess item characteristics, such as comprehensibility and unambiguity

Differentiation between criterion (or contrast) groups: Test size and direction of expected differences between groups on the construct of interest

Factor analysis: Test hypothesized structure of the construct domain (e.g., whether items thought to define the construct load on the same factor and not on other factors)

Correlation: Test relation between measure of construct and measure of other distinct constructs

Multitrait multimethod: Test whether different measures of the same construct correlate more highly than measures of different constructs using same and different methods (e.g., instruments, data sources, languages)

Criterion-oriented (or external) validity had traditionally been considered most central because a test or measure that fails to predict or relate to anything else of interest would be of little use. At the time, it seemed useful to distinguish concurrent validity (the extent to which the test relates to relevant criteria

obtained at the same time) and predictive validity (the extent to which the test can predict relevant variables, events, and outcomes in the future).

The “chief innovation” (Cronbach & Meehl, 1955, p. 281), however, was the notion of construct validity. Many researchers had come to appreciate the so-called criterion problem – that any one external criterion is also only a “measure” that is itself an imperfect indicator of the construct to be measured and thus cannot, by itself, truly and fully represent the construct. If there is no gold standard, no single criterion against which the test can be validated, how should we establish inferences about the proper interpretation or meaning of the scores on the test? The question thus became: “What constructs account for variance in test performance?” (Cronbach & Meehl, 1955, p. 282).

An Integrated Conception of Construct Validity

It was soon recognized that the early view of several distinct types of validity was fragmented and misguided. Loevinger (1957) extended the theoretical implications of construct validity by proposing that all scientific issues in test construction, validation, and test use be evaluated from the construct point of view. A *construct* is a hypothetical attribute, process, or other regularity in the behavior of individuals, groups, or other entities (e.g., liberal values, extraversion, self-monitoring tendencies), and procedures for determining the validity of a measure are similar to the general scientific procedures for developing and confirming theories. That is, what seemed like different types of validity are really just different sources of evidence that address particular questions of construct validity.

It is now generally agreed that the construct validity of our observed variables is the central concern in measurement. This is quite a departure from classical test theory, which holds that measures are imperfect indicators of constructs because they contain some degree of random measurement error or unreliability. The construct view, in contrast, argues that the variables we observe or measure are imperfect indicators not only because of random errors; more important, they are imperfect because they also measure constructs we did not intend to measure and thus include *systematic* error (e.g., error introduced by the particular method used to collect the data). In this view (e.g., Judd & McClelland, 1998), scores on observed variables potentially reflect three sources of variance: (1) the construct we intend to measure (convergent aspects of validation), (2) a variety of other constructs (or sources of influence) we would like to avoid measuring (discriminant aspects of validation), and (3) random error (or unreliability). This

broad construct view thus highlights convergent and discriminant validity and considers reliability as just another piece of evidence for the construct validity of the proposed measurement.

Messick (1989, 1995) has articulated a comprehensive program of construct validation that addresses the meaning of test scores in test interpretation and use. His view highlights that validity is an “integrative evaluative judgment of the degree to which evidence and theoretical rationales support the *adequacy* and *appropriateness*” of the theoretical specification of the construct (Messick, 1989, p. 13). Thus, validity is considered a property of the interpretation of a measure rather than a property of the measure itself; for example, there may be substantial evidence to support the interpretation of a particular attitude scale as a measure of individual differences in liberal values but no validity evidence for its interpretation as a measure of intelligence or extraversion. Of course, if the theoretical account of the construct is specified clearly and in detail, specific predictions about relations to other constructs and criteria can be readily made, thus simplifying the process of collecting evidence that supports or disconfirms a particular interpretation of the test score.

Like any other theory or model, the validity of the particular score interpretation can never be established but is always evolving to form an ever-growing “nomological network” of validity-supporting relations (Wiggins, 1973). Given that multiple pieces of evidence will accumulate to support the hypothesized construct, it is often difficult to summarize the available validity evidence with a simple quantitative index, and investigators have had to resort to qualitative and tabular summaries. For example, Snyder (1987) wrote a whole book to summarize what has been learned about the self-monitoring construct in more than 20 years of empirical research and construct development. More recently, meta-analytic techniques (see Johnson & Eagly, Chapter 26 in this volume) have proven useful to make such data summaries more manageable and objective (Schmidt, Hunter, Pearlman, & Hirsch, 1985).

Types of Evidence for Construct Validity

In his integrative account, Messick (1989) specified six forms of evidence that should be sought to examine construct validity. The six forms of evidence are listed and defined briefly in the middle section of [Table 18.4](#).

The first is evidence for *content* validity; such evidence is provided most easily if the construct has been explicated theoretically in terms of specific aspects that exhaust the content domain to be covered by the construct. Common

problems involve underrepresenting an important aspect of the construct definition in the item pool and overrepresenting another one. An obvious example are the multiple-choice exams we often construct to measure student performance in our classes; if the exam questions do not sample fairly from the relevant textbook and lecture material, we cannot claim that the exam validly represented what students were supposed to learn (i.e., the course content).

Arguments about content validity arise not only between professors and students, but also in research. For example, when revising his self-monitoring scale, Snyder (1987) excluded a number of items measuring other-directed self-presentation, thus representing behavioral variability and attitude-behavior inconsistency to a lesser extent in the revised scale; because all items measuring public performing skills were retained, the construct definition in the new scale shifted toward a conceptually unrelated construct, extraversion (John, Cheek, & Klohn, 1996). This example shows that discriminant aspects are also important in content validation; to the extent that the items measure aspects not included in the construct definition, the measure would be contaminated by construct-irrelevant variance. For example, when validating scales to measure positive and negative emotion expression, these scales should not assess variance that must be attributed to theoretically unrelated constructs, such as social desirability or self-esteem (e.g., Gross & John, 1997).

As shown in the third section of Table 18.4, there are a number of validation procedures researchers might use (see also Smith & McCarthy, 1995). Researchers might ask expert judges to review the match between item representation and construct domain specification, and to add or delete items. Another procedure would be to use factor analysis to verify the hypothesized structure of the content domain.

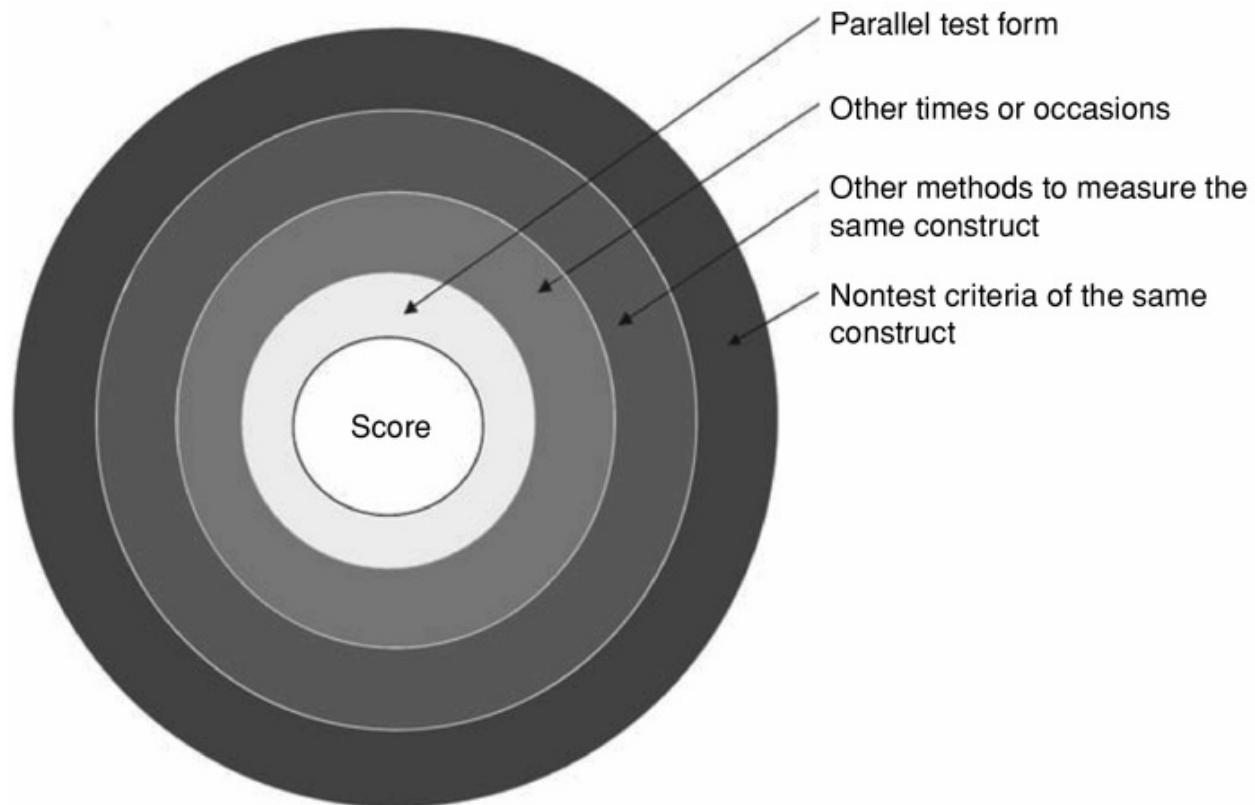


Figure 18.3. How far can we generalize from a test score? The onion model of generalizability.

Substantive validity evidence makes use of substantive theories and process models to further support the interpretation of the test scores. Relevant procedures might involve differentiation between criterion (or contrast) groups assumed to differ in the relevant processes. For example, Cacioppo and Petty (1982) developed the Need for Cognition Scale to measure individual differences in the preference and enjoyment of effortful thinking; as part of their construct validation program, they conducted a study contrasting college professors (assumed to need cognition) and assembly line workers (assumed to not need cognition). Even stronger evidence for substantive validity comes from studies that use experimental manipulations that directly vary the processes in question. For example, Petty and Cacioppo (1986) showed that the process of attitude change was moderated by the need for cognition: Individuals scoring high on the scale were influenced by careful examination of the arguments presented in a message, whereas those scoring low were more influenced by extraneous aspects of the context or message (e.g., the attractiveness of the source of the message).

Structural validity evidence requires that the correlational (or factor) structure of the measure is consistent with the hypothesized internal structure of the construct domain. We noted the issue of multidimensionality in the section on reliability, pointing out that coefficient alpha does not allow inferences about the dimensionality of a measure. The structure underlying a measure or scale is not an aspect of reliability; rather, it is central to the interpretation of the resulting scores and thus needs to be addressed as part of the construct validation program. Researchers have used both exploratory and confirmatory factor analysis for this purpose, and we will return to this important issue later in the chapter in the context of evaluating measurement with structural equations models.

Generalizability evidence, as used here by Messick (1989, 1995), is needed to demonstrate that score interpretations apply across tasks or contexts, times or occasions, and observers or raters. The inclusion of generalizability evidence here makes explicit that construct validation includes consideration of “error associated with the sampling of tasks, occasions, and scorers (that) underlie traditional reliability concerns” (Messick, 1995, p. 746). In this context we should note that the notion of generalizability encompasses traditional conceptions of both reliability and criterion validity; they may be considered on a continuum, differing only in how far generalizability claims can be extended (Thorndike, 1997). Traditional reliability studies provide relatively “weak” tests of generalizability, whereas studies of criterion validity provide “stronger” tests of generalizability.

As suggested by Figure 18.3, generalizing from a test score to another test constructed according to parallel procedures increases our confidence in the test, but does so only modestly. If we find we can also generalize to other times or occasions, our confidence is further strengthened, but not by quite as much as when we can show generalizability to other methods or even to nontest criteria related to the construct the test was intended to measure. Figure 18.3 thus resembles the layers of an onion, showing how far the test allows us to generalize, with the inner layers representing relatively modest levels of generalization and the outer layers representing farther-reaching generalizations to contexts that are more and more removed from the central core (i.e., dissimilar from the initial measurement operation).

The kind of validity evidence Messick (1989) considered under the generalizability rubric is crucial for establishing the limits or boundaries beyond which the interpretation of the measure cannot be extended. An issue of

particular importance for social-personality researchers is the degree to which findings generalize from “convenience” samples, such as American college students, to groups that are less educated, older, or come from different ethnic or cultural backgrounds.

Consequential validity evidence focuses on the personal and societal consequences (both intended and unintended) of score interpretation and use. It requires the test-user to confront issues of test bias and fairness and is of paramount importance in contexts where tests are used to make important decisions about individuals. Thus, it is more about valid use of the test than about the validation of the test per se. Consequential validity is a greater concern in educational and employment settings than in social-personality research contexts, where scale scores and performances in experimental tasks have little, if any, consequence for the research participant. Finally, *external* validity covers such a broad range of both convergent and discriminant evidence that we consider it in more detail.

External Validation: Convergent and Discriminant Aspects

External validity evidence refers to the ability of a test to predict conceptually related behaviors, outcomes, or criteria, and has been emphasized by a wide range of writers. For example, in their review of attitude measurement, Dawes and Smith (1985, p. 512) argued that “the basis of all measurement is empirical prediction,” and in his review of personality measurement, Wiggins (1973, p. 406) argued that prediction “is the sine qua non of personality assessment.” Obviously, it makes sense that a test or scale should predict construct-relevant criteria. It is less apparent that we also need to show that the test does not predict conceptually unrelated criteria. In other words, a full demonstration of external aspects of construct validation requires a demonstration of both what the test measures and what it does not measure.

Multitrait Multimethod Matrix. Campbell and Fiske (1959) introduced the terms “convergent” and “discriminant” to distinguish demonstrations of what a test measures from demonstrations of what it does not measure. The *convergent* validity of a self-report scale of need for cognition could be assessed by correlating the scale with independently obtained peer ratings of the participant's need for cognition and with frequency of effortful thinking measured by beeping the participant several times during the day. *Discriminant* validity could be assessed by correlating the self-report scale with peer ratings of extraversion and

a beeper-based measure of social and sports activities. Campbell and Fiske were the first to formalize these ideas of convergent and discriminant validity into a single systematic design that crosses multiple traits or constructs (e.g., need for cognition and extraversion) with multiple methods (e.g., self-report, peer ratings, and beeper methodology). They called this design a multitrait multimethod (MTMM) matrix, and the logic of the MTMM is both intuitive and compelling.

What would we expect for our example? Certainly, we would expect sizable convergent validity correlations among the need for cognition measures across the three methods (self-report, peer report, beeper); because these correlations involve the same trait but different methods, Campbell and Fiske (1959) called them monotrait-heteromethod coefficients. Moreover, given that the need for cognition is theoretically unrelated to extraversion, we would expect small discriminant correlations between the need-for-cognition measures and the extraversion measures; this condition should hold even if both traits are measured with the same method, leading to so-called heterotrait-monomethod correlations. Certainly, we want each of the convergent correlations to be substantially higher than the discriminant correlations involving the same trait. And finally, the same patterns of intercorrelations among the constructs should emerge, regardless of the method used; in other words, the relations among the constructs should generalize across methods.

Method Variance. One important recognition inherent in the MTMM is that we can never measure the trait or construct by itself; rather, we measure the trait intertwined with the method used: “[E]ach measure is a trait-method unit in which the observed variance is a combined function of variance due to the construct being measured and the method used to measure that construct” (Rezmovic & Rezmovic, 1981, p. 61). The design of the MTMM is so useful because it allows us to estimate variance in our scores that is attributable to method effects – that is, errors systematically related to our measurement methods and thus conceptually quite different from the notion of random error in classical test theory. These errors are systematic because they reflect the influence of unintended constructs on scores, that is, unwanted variance – something we did not wish to measure but that is confounding our measurement (Ozer, 1989).

Method variance is indicated when two constructs measured with the same method (e.g., self-reported attitudes and self-reported behavior) correlate more highly than when the same constructs are measured with different methods (e.g., self-reported attitudes and behavior coded from videotape). For example, it has

been argued that positivity bias in self-perceptions is psychologically healthy (Taylor & Brown, 1988); however, if positivity bias is measured with self-reports and the measure of psychological health is a self-report measure of self-esteem, then the positive intercorrelation between these measures may not represent a valid hypothesis about the two constructs (positivity bias and psychological health) but shared self-report method variance associated with narcissism (John & Robins, 1994); that is, individuals who see themselves too positively may be narcissistic and also rate their self-esteem too highly. Discriminant validity evidence is needed to rule out this alternative hypothesis, and the construct validity of the positivity bias measure would be strengthened considerably if psychological health were measured with a method other than self-report, such as ratings by clinically trained observers (Robins & John, 1997).

Multiple Sources of Data: LOTS. Beginning with Cattell (1957, 1972), psychologists have tried to classify the many sources researchers can use to collect data into a few broad categories. Because each data source has unique strengths and limitations, the construct validation approach emphasizes that we should collect data from *lots* of different sources, and so the acronym LOTS has particular appeal (Block & Block, 1980).

L data refer to life-event data that can be obtained fairly objectively from the individual's life history or life record, such as graduating from college, getting married or divorced, moving, socioeconomic status, memberships in clubs and organizations, and so on. Examples of particularly ingenious measures derived from *L* data are counts of bottles and cans in garbage containers to measure alcohol consumption (Webb, Campbell, Schwartz, Sechrest, & Grove, 1981) and police records of arrests and convictions to measure juvenile delinquency (Moffitt, 1993).

O data refer to observational data, ranging from observations of very specific aspects of behavior to more global ratings (see Heyman, Lorber, Eddy & West, Chapter 14 in this volume; Kerr, & Tindale, Chapter 9 in this volume). Examples are careful and systematic observations recorded by human judges, such as in laboratory settings or carefully defined situations; behavior coded or rated from videotape; and reports from knowledgeable informants such as peers, roommates, spouses, teachers, and interviewers that may aggregate information across a broad range of relevant situations in the individual's daily life. *O* data obtained through unobtrusive observations or coded later from videotape can be particularly useful to make inferences about the individual's attitudes, prejudices,

preferences, emotions, and other attributes of interest to social scientists. A nice illustration is a study that recorded seating position relative to an out-group member to measure ethnocentrism (Macrae, Bodenhausen, Milne, & Jetten, 1994).

T data refer to information from test situations that provide standardized measures of performance, motivation, or achievement, and from experimental procedures that have clear and objective rules for scoring performance. Reaction times are frequently used in studies of social cognition, providing an objective measure of an aspect of performance. An intelligence test is another kind of example. A third is the length of time an individual persists on a puzzle or delays gratification in a standardized situation (Mischel, 1990).

Last, but not least, *S* data refer to self-reports. *S* data may take various forms. Global self-ratings of general characteristics and true-false responses to questionnaire items have been used most frequently. However, self-reports are also studied in detailed interviews, in narratives and life stories, and in survey research (Visser, Krosnick, Lavrakas, & Kim, Chapter 16 in this volume). Daily experience sampling procedures (see Reis, Gable & Maniaci, Chapter 15 in this volume) can provide very specific and detailed self-reports of moment-to-moment functioning in particular situations.

The logic underlying *S* data is that individuals are in a good position to report about their psychological processes and characteristics – unlike an outside observer, they have access to their private thoughts and experiences and can observe themselves over time and across situations. However, the validity of self-reports depends on the ability and willingness of the individual to provide valid reports, and self-reports may be influenced by various constructs other than the intended one. Systematic errors include, most obviously, individual differences in response or rating scale use, such as acquiescence (see Visser, Krosnick, Lavrakas, & Kim, Chapter 16 in this volume) and response extremeness (Hamilton, 1968).

Moreover, some theorists have argued that self-reports are of limited usefulness because they may be biased by social desirability response tendencies. Two kinds of desirability biases have been studied extensively (for a review, see Paulhus & John, 1998). Impression management refers to deliberate attempts to misrepresent one's characteristics (e.g., “faking good”), whereas self-deceptive enhancement reflects honestly held but unrealistic self-views. Impression management appears to have little effect in research contexts where individuals participate anonymously and are not motivated to present themselves

in a positive light; self-deception is not simply a response style but is related to substantive personality characteristics, such as narcissism.

Fortunately, although social-personality psychologists use self-reports most frequently, other methods are available and used. Thus, measures based on *L*, *O*, and *T* data can help evaluate and provide evidence for the validity of more easily and commonly obtained self-report measures tapping the same construct. Unfortunately, research using multiple methods to measure the same construct has not been very frequent. Overall, it seems that multimethod designs have been underused in construct validation efforts. In a way, researchers seem more likely to talk about the MTMM approach than to go to the trouble of actually using it.

There is an extensive and useful methodological literature on the MTMM, which took off in the 1970s when SEM became available and provided powerful analytical tools to estimate separate trait and method factors (Kenny, 1976; Kenny & Kashy, 1992; Schwarzer, 1986; see also Fabrigar & Wegener, Chapter 19 in this volume). A number of excellent reviews and overviews have appeared recently. For example, West and Finch (1997, pp. 155–159) provided hypothetical data to illustrate three scenarios: (1) convergent and discriminant validity with minimal method effects, (2) strong method effects, and (3) effects of unreliability and lack of discriminant validity. Judd and McClelland (1998, tables 13.11–13.15) provide a series of examples that illustrate Campbell and Fiske's (1959) original principles of convergent and discriminant validation as well as the application of SEM techniques to estimate separate trait and method effects.

To summarize, in this section we reviewed Messick's (1989) six forms of evidence relevant to construct validation (see Table 18.4) and then considered one of them, external validation, in some detail, focusing on convergent and discriminant aspects such as the multitrait-multimethod approach, the nature of method variance, and multiple sources of data. Although one might quibble with some of Messick's particular categories (e.g., some of them seem to overlap), we view his formulation as comprehensive and heuristically useful. Most important, we agree with Messick's view that evidence concerning traditional issues of reliability are part of the construct validation program, namely under the heading of generalizability, and that evidence about dimensionality must be considered in the context of structural validity. In the following section we reconsider these issues, now from the perspective of the measurement model in SEM.

Model Testing in Construct Validation and Scale Construction

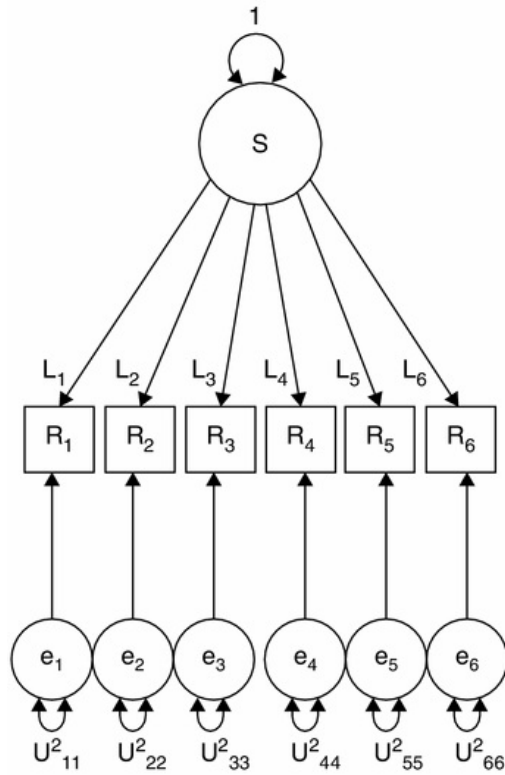
The measurement model in SEM is based on confirmatory factor analysis (CFA) (see Bollen & Long, 1993; Fabrigar & Wegener, Chapter 19 in this volume; Loehlin, 1998, McArdle, 1996). CFA is particularly valuable because it provides a general analytic approach to assessing construct validity. As will become clear, convergent validity, discriminant validity, and random error can all be addressed within the same general framework. To illustrate these points, we briefly discuss a simple numerical example.

Measurement Models in SEM: Convergent Validity, Discriminant Validity, and Random Error

Like all factor analytic procedures (Floyd & Widaman, 1995; Tinsley & Tinsley, 1987; Widaman & Grimm, Chapter 20 in this volume), CFA assumes that a large number of observations or items are a direct result (or expression) of a smaller number of *latent* sources (i.e., unobserved, hypothetical, or inferred constructs). However, CFA eliminates some of the arbitrary features often criticized in exploratory factor analysis (Gould, 1981; Sternberg, 1985). First, CFA techniques require the researcher to specify an explicit model (or several competing models) of how the observed (or measured) variables are related to the hypothesized latent factors. Second, CFA offers advanced statistical techniques that allow the researcher to test how well the a priori model fits the particular data; even more important, CFA permits comparative model testing to establish whether the a priori model fits the data better (or worse) than plausible alternative or competing models.

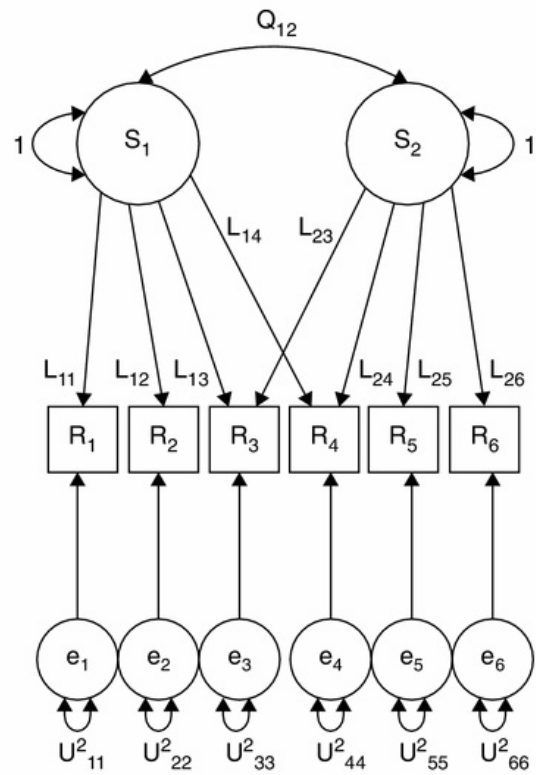
CFA models can be displayed graphically, allowing us to effectively communicate the various assumptions incorporated in each model. Some examples are shown in Figure 18.4. Figure 18.4a shows a common-factor model, in which a single underlying construct S (shown as a circle on the top) is assumed to give rise to the correlations among the six items or responses $R1$ to $R6$ (the observed variables, shown in squares). Following established convention (Bentler, 1980), circles are used to represent latent variables, whereas squares represent measured (or manifest) variables; arrows with one head represent directed or regression parameters, whereas two-headed arrows (which are often omitted) represent covariance of undirected parameters. Note that each measured variable R_m has two arrows leading to it. The arrow from the latent construct S is

a factor loading L_m that represents the strength of the effect that the latent construct has on each observed variable. The other arrow involves another latent variable for each observed variable – these are unique factor scores (em) that represent the unique or residual variance (U^2) remaining in each observed variable.



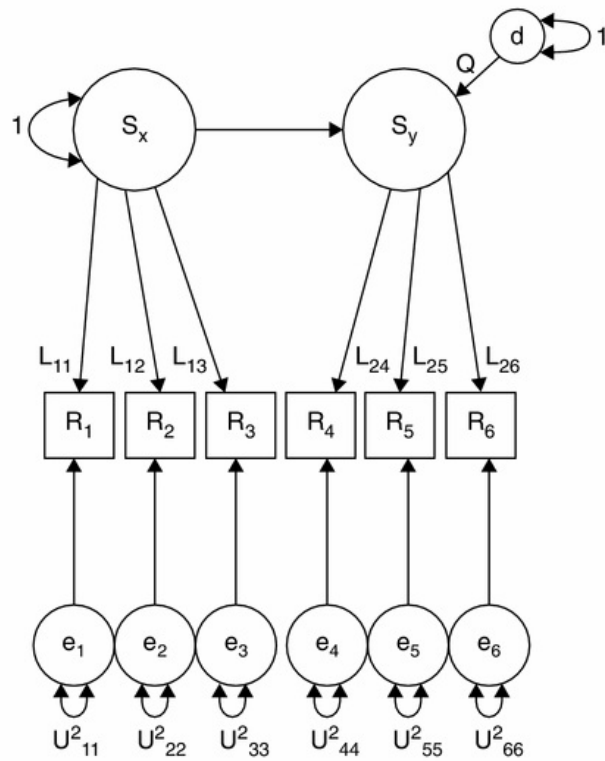
One Common Factor Model

(a)



Two Factor Oblique Model

(b)



Regression Model Between Factors

(c)

Figure 18.4. Measurement models in structural equation modeling: (a) a one-common-factor model, (b) a two-factor oblique model, and (c) a model showing one common factor related to a criterion.

Conceptually, this model captures a rather strong structural hypothesis, namely that the six observed variables covary only because they all measure the same underlying construct *S*. In other words, we hypothesize that the only thing the items have in common is this latent construct, and all remaining or residual item variance is idiosyncratic to each item and thus unshared. This structural model provides a new perspective on how to define two important terms we have used in this chapter: the convergent validity of the item and random error. In particular, the loading of an item on the construct of interest represents the convergent validity of the item, whereas its unique variance represents random error. However, in this simple measurement model, we cannot address discriminant validity.

Compare the measurement model in Panel a of [Figure 18.4](#) with the one in Panel b that postulates two factors *S1* and *S2* influencing responses to six items. Here we are hypothesizing two distinct constructs, rather than one. Note that this model incorporates another condition, known as *simple structure*: The convergent validity loadings (represented by arrows from the latent constructs to the observed items) indicate that the first two items are influenced by the first construct but not the second construct, whereas the last two items are influenced only by the second construct and not the first. In other words, these items can be uniquely assigned to only one construct, which much simplifies the measurement model. With two constructs in the measurement model, we can also address issues of discriminant validity. Whereas the item's loading on the construct of interest represents convergent validity and its unique variance random error, its loading on constructs other than the intended one is relevant to discriminant validity.

Note that this model includes an arrow between the two constructs, indicating a correlation or covariance; the two constructs are not independent (orthogonal) but rather related (oblique). At the level of the constructs, this correlation tells us about discriminant validity. If the correlation is very high (e.g., .90), we would worry that the two constructs are not distinguishable and that we really have only one construct; if the correlation is very low (e.g., .10), we would be reassured that the two concepts show good discriminant validity with respect to

each other. There is another possibility here, namely that the two constructs could be components of a broader, superordinate construct that includes them both. These issues involve questions about the dimensionality and internal structure of the constructs being measured. We discussed these issues earlier in the section on reliability but, as we argued in the section on validity, dimensionality issues are part of the construct validation program (see [Table 18.4](#)) because they concern the structural validity of the interpretation of our measure.

Structural Validity Examined with SEM: An Empirical Example. Structural validity issues resurface with great regularity in the social-personality literature. Some of the most popular constructs have endured protracted debates about their validity: self-monitoring, attributional style, hardiness, Type A coronary-prone behavior pattern, and need for closure (Hull, Lehn, & Tedlie, [1991](#); Neuberg, Judice, & West, [1997](#)). Part of the problem is that many of these constructs, and the scales designed to measure them, were initially assumed to be unidimensional, but later evidence challenged those initial assumptions. It is therefore instructive to consider how SEM approaches can help address the underlying issues and to provide a numerical though manageable example as an illustration.

For the purpose of this illustration we constructed two hypothetical scales and then used actual data from participants who had rated themselves on a number of personality-descriptive adjectives and phrases (Benet-Martínez & John, [1998](#)); we used a large sample ($N = 450$), because small sample sizes can create problems for SEM estimation procedures (McArdle, [1996](#)).

The first scale was intended as a measure of impulsivity (vs. inhibition), a construct of long-standing interest and debate (e.g., Block, [1995](#); Kagan & Snidman, [1991](#)). To address content validation early on, we defined our universe of item content from the perspective of generalizability theory, using a design that varied two facets of generalizability: context (task vs. social) and construct pole (impulsive vs. inhibited). For the high (impulsive) pole, we selected from our existing item pool three items to represent task contexts (careless, disorganized, and lazy) and two items to represent social contexts (enthusiastic and assertive); for the low (inhibited) pole of the construct, two items each were selected for task contexts (persevering and thorough) and social contexts (reserved and shy).

The second scale, briefly, was intended as a measure of extraversion and sampled from previously studied content facets of the construct, namely

talkativeness and self-assertion; again, it included both extraverted items (assertive, has an assertive personality, bold, verbal, is talkative, and talkative) and introverted items (untalkative, tends to be quiet, and timid). To begin with, we assumed that each scale is unidimensional (of course, we had doubts about one of the scales, as will soon become clear).

What do we find when we apply the traditional analyses of internal consistency and exploratory factor analysis to these two scales? [Table 18.5](#) summarizes the major results. As we cautioned earlier, one should not calculate alphas before the unidimensionality of a scale has been verified. Indeed, the alphas seem reasonably high for these relatively short scales, and most journal editors would consider even the lower alpha of .79 quite acceptable (in fact, the 25-item self-monitoring scale had a lower alpha than this 9-item scale; Snyder, 1987). Moreover, the item-total correlations were all substantial for all items on both scales.

Table 18.5. *Structural Validity Example: Traditional Psychometric Characteristics of the “Impulsivity” and Extraversion Item Sets*

	“Impulsivity”	Extraversion
	(two-dimensional)	(one-dimensional)
Alpha of total scale	79	90
First principal component	38%	58%
Second principal component	20%	13%
Correlation between subscales	30	67

Note: N = 452 college students. Data from Benet-Martínez and John (1998).

What about the structural validity of these scales? Exploratory factor analyses

resulted in eigenvalues (Table 18.5 shows the corresponding percentages of variance accounted for by the first and second unrotated principal component) that make it hard to tell conclusively if we need one or two factors, although for the impulsivity scale the evidence points to two factors. The loadings for two rotated factors, shown in Figure 18.5a and 18.5b, are more informative. For impulsivity, the items defining our four a priori facets nicely hang together, and the items show excellent simple structure. However, the exploratory factor solution seems inconsistent with a one-dimensional impulsivity theory because the social impulsivity items and the task impulsivity items formed two distinct factors.

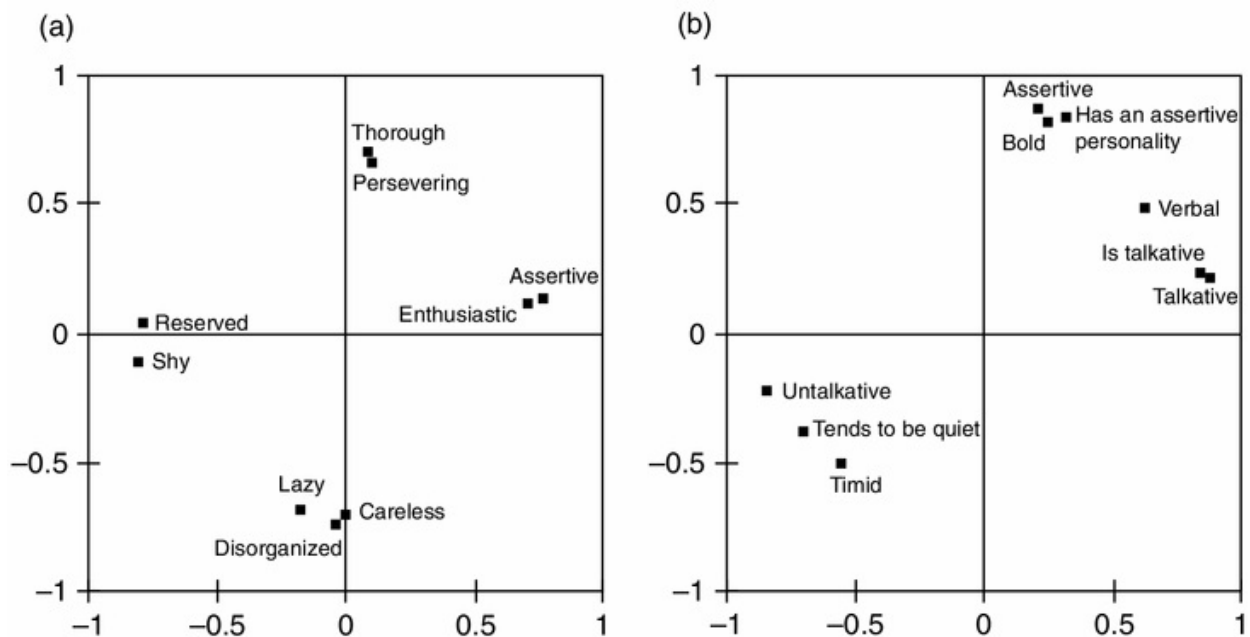


Figure 18.5. Plot of exploratory factor loadings (after Varimax rotation) for (a) the items assumed to relate to Impulsivity and (b) items assumed to relate to Extraversion.

To find out whether the social and task impulsivity factors have a substantial positive relation, as the general impulsivity account would suggest, we formed two scales and intercorrelated them. Social impulsivity (enthusiastic and assertive vs. reserved and shy) correlated $-.30$ with task impulsivity (careless, lazy, and disorganized vs. thorough and persevering), thus failing to produce the predicted positive correlation – the facets do not hang together the way they should.

In contrast, if we rotate two factors for the extraversion scale (see Figure 18.5b), we find much less evidence for simple structure; instead, the variables all

fall into a positive manifold (all items are either in the high-high or low-low quadrant), with the two a priori facets forming somewhat separable assertiveness and talkativeness clusters, especially within the upper-right quadrant. As expected from this loading pattern, the two clusters (when scored as scales) were correlated .67 – a positive and substantial correlation (see [Table 18.4](#)). These exploratory analyses leave us with some alternative hypotheses that we can test against the a priori models, using SEM. The SEM analyses are summarized very briefly in [Table 18.6](#). We begin with the one-factor model because it is the simplest or “compact model” (Judd et al., 1995). Because the models are all nested, we can statistically compare them with each other, testing the relative merits of more complex (i.e., full or augmented) models later. Without going into detail, the comparative fit indices show that we can clearly reject the one-factor model for the impulsivity scale; a test comparing it with the uncorrelated two-factor model shows that the latter provides a significantly better fit for the data. Not surprisingly, fit is improved further (and significantly) when correlations between the two factors are freely estimated. Note, however, that this model requires the estimation of another parameter, namely the $-.30$ correlation between the two factors.

These findings are inconsistent with the general impulsivity hypothesis: Because the two factors do not correlate positively, we cannot conclude that the scale consists of two impulsivity facets that together form the superordinate construct. Rather, we might conjecture that we are measuring abbreviated versions of the familiar Big Five factors of extraversion (defined by our social impulsivity items) and conscientiousness (defined by the task impulsivity items reversed-scored). Indeed, the present findings are quite consistent with numerous studies of much broader sets of personality descriptors (e.g., John, 1990; John & Srivastava, 1999).

Criteria for Unidimensionality. We can push the model comparison approach even farther. Rather than letting SEM estimate the correlation (or covariation) between the latent factors as a free parameter, we can fix it at a value that would allow us to make strong inferences about the independence of the two constructs. This is what we did in the last two models in [Table 18.6](#): We set a specific decision rule, fixing the correlation to one of two decisive values and asking which one fit the data more closely.

Hattie (1985) emphasized that researchers need to formulate decision criteria that help them decide “how close a set of items is to being a unidimensional set” (p. 159). In most real data, just as in the present one, unidimensionality is not

simply present or absent. Thus, we need to make a conceptual argument at what levels of intercorrelation we will call a measure relatively unidimensional or relatively multidimensional. Although reasonable people can disagree about any one cut-off point because it is inherently arbitrary, we are prepared to argue that factor intercorrelations as low as .20 would indicate relative independence, whereas correlations as high as .80 suggest such substantial overlap that a one-factor model should be preferred on the basis of parsimony.

As shown in Table 18.6, this decision rule allowed us to differentiate between the two models. The low-intercorrelation model provided a significantly better fit for the “impulsivity” items than did the high-intercorrelation model, thus correctly identifying this item set as measuring two essentially independent constructs; we say correctly here because we had in fact constructed this set by drawing items from uncorrelated Big Five self-report scales for conscientiousness and extraversion, respectively. In contrast, note that the high-intercorrelation model provided a significantly better fit for the extraversion scale. This result suggests that this item set is best interpreted as measuring an essentially unidimensional construct with two highly correlated item clusters, which might be interpreted as talkative and assertive manifestations of extraversion.

Table 18.6. Structural Validity Example in the SEM Approach: Summary of Models Tested and Their Fit Indices

Model Tested	df	Comparative Fit Index	
		“Impulsivity”	Extraversion
One factor only	27	.57	.82
Two uncorrelated factors	27	.90	.83
Two correlated factors,	26	.93	.93
<i>r</i> freely estimated		(<i>r</i> = −.30)	(<i>r</i> = .71)
Two correlated factors,	27	.92	.87
<i>r</i> set to .20			

More Complex Models Including External Validity. In a fully developed construct validation program, of course, we would not stop here. For the “impulsivity” item set, we would move on to testing the relations of these two SEM-based constructs with other measures of extraversion and conscientiousness, preferably drawn from other data sources, such as peer ratings or behavioral observations. Using an MTMM design to address external validity, we would gather evidence both about convergent validity (e.g., self-reported conscientiousness with measures of conscientiousness drawn from other data sources) and discriminant validity (e.g., measures of conscientiousness with measures of extraversion from the same data source).

Again, we would use SEM procedures for these additional validation steps, as suggested by the simplified model in [Figure 18.4c](#). This model shows how we can represent a unidimensional measurement model for construct S_x and a unidimensional criterion construct S_y , along with a predictive (or convergent) validity relation represented by the arrow from S_x to S_y . Note that this model addresses the criterion problem that seemed so intractable in the early treatments of validity: The criterion itself is not treated as a gold standard but modeled as a construct that must also be measured with fallible observed indicator variables. We should note that the models used to represent trait and method effects in MTMM matrices are considerably more complex than are the simple models considered here; for example, McArdle (1996, figure 13.2) provided an elegant model for a more complete representation of the construct validation program.

Many readers might benefit from an example with more extensive numerical illustrations of SEM than we could provide here. We recommend an early paper by Judd, Jessor, and Donovan (1986), who examined the construct validity of a nine-item scale designed to measure attitudinal tolerance of deviance, including attitudes toward shoplifting, lying, and getting into fights. This construct postulates the existence of an underlying general attitude toward deviance

manifested in self-reported attitudes about specific deviant behaviors. To elaborate four aspects of the construct validity of this measure, Judd *et al.* (1986) used various SEM procedures. First, to examine the convergent validity (or internal consistency) of the nine items, they analyzed their intercorrelations (or covariances), testing hypotheses about structural validity (e.g., do all nine items reflect a single common factor?). Second, to examine external (or criterion) validity, they tested whether the construct relates to other constructs in theoretically consistent ways (e.g., do these attitudinal items predict deviant behavior?). Third, to address discriminant validity, they measured discriminant relations regarding religious attitudes in terms of both structural validity (e.g., are the attitude-toward-deviance items reliably different from religious attitude items?) and criterion validity (e.g., are these items better at predicting deviant behaviors than are the religious attitude items?). Fourth, they investigated particular aspects of substantive validity, namely temporal stability and prediction of behavior over time; these substantive predictions are important because attitudes, like other personality constructs, refer to individual differences that are assumed to be relatively stable and enduring over time, rather than transitory states of short duration (Chaplin, John, & Goldberg, 1988).

Issues in Questionnaire Construction

So far, we have discussed construct validation as if the measure to be validated already existed. However, construct validation issues are central not only during the evaluation of existing psychological measures but also during each stage of their development. We now consider the somewhat specialized case of questionnaire (or scale) construction. The first questionnaires were developed in the early 1900s, and since the 1950s the construction of questionnaires began to proliferate (Goldberg, 1971). We argue that questionnaire construction, like the development of any psychological measure, must be considered in the context of a program of construct validation. Historically, however, the construct validation approach was not articulated until the 1950s, and the consensus in its favor has been building slowly and quietly, mostly since the 1970s, and it is far from complete. Three distinct schools of thought preceded it and retain adherents even today.

Early Approaches: External, Rational-Intuitive, and Internal. Three approaches to questionnaire construction emerged in the 1950s; each was inspired by one particular type of validity (see Table 18.4) and aimed to maximize that particular type of validity while ignoring all others (for reviews,

see Burisch, 1984, 1986). Given today's perspective favoring an integrated construct approach, the ideological fervor of these three camps strikes us almost like self-parodies.

The so-called *external* approach emphasized maximizing criterion validity, seemingly lost in “a single-minded bivariate search for items that correlate with a chosen criterion” (Tellegen, 1985, p. 685). Typically, externally oriented researchers would administer large and atheoretically assembled sets of questionnaire items to preselected criterion and control groups (e.g., patients hospitalized for depression versus patients admitted for surgical procedures) and then determine empirically which items significantly differentiated the two groups. The items that successfully differentiated between the groups would be retained to form the resulting scale (e.g., for depression), regardless of the actual item content or broader theoretical considerations. The most famous products of the external approach are the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943) for clinical populations and the California Psychological Inventory (CPI; Gough, 1987) for normally functioning adults. The continued popularity of these instruments, conceived in the 1950s, is testimony to the endurance of the approach. Although the obsession with criterion validity still persists in some literatures (e.g., on marital interaction and satisfaction), the external approach largely fell out of favor, primarily because the subtle and theoretically opaque items did not form psychologically coherent and heuristic constructs, were hard to replicate, and required a rather inefficient scale-construction process.

At the other extreme of the dust bowl empiricists were those psychologists who had detailed theories they did not doubt. Thus they felt free to focus solely on the content and face validity of their measures. Variouslly labeled the *rational*, *intuitive*, or *deductive* approach, they easily generated items on the basis of their theories. The resulting scales, face-valid with obvious item content, proved remarkably popular, if not always with other researchers then certainly with the test-taking public. In fact, this approach gave birth to the Myers-Briggs Type Indicator (MBTI; Myers & McCaulley, 1985), based partly on Carl Jung's type theories. Without much evidence for its external, structural, or substantive validity, the MBTI nonetheless became the most popular personality questionnaire in this country. To the eternal embarrassment of research psychologists, the MBTI continues to be used at major research universities in applied contexts, such as counseling and career advising. On the brighter side, the deductive approach eventually developed into the construct approach, which, as we have described earlier, is more interested in empirical evidence that might

turn out to disconfirm one's favorite theory.

Finally, an emphasis on structural validity and the growing availability of exploratory factor analysis in the 1950s and 1960s gave rise to the *internal* or *inductive* approach to questionnaire construction. As the label suggests, the focus was on discovering the factor structure of large-item sets, often assembled with little concern for particular content representation or selection. The early factor-analytic personality models of Cattell, Eysenck, and Guilford were based on this approach and dominated until the mid-1980s when they gave way to the emerging consensus on the Big Five dimensions (Goldberg, 1993; John, 1990; John & Srivastava, 1999). The preoccupation with the internal factor structure in self-reports came at the expense of other sources of validity evidence. Partly because the dimensions emerging from factor analyses were assumed to be “real,” theoretical construct definitions, substantive validity evidence, and criterion validation against measures of behaviors were deemed of secondary interest.

It is easily apparent that each of these three approaches, in its pure form, had a great strength that was also its greatest failure, namely its single-minded pursuit of just one type of validity evidence. Obviously some kind of integration was needed. Although the conceptual foundations had been laid already in the 1950s, the construct validation approach emerged only gradually, as the three earlier approaches grew softer around the edges and eventually became indistinguishable.

Recapitulation: Modern Construct-Oriented Scale Construction. Few, if any, scales or measures today are constructed according to just one of the early approaches. Most researchers have adopted, implicitly or explicitly, many of the features of the construct validation program discussed in this chapter. In fact, much of our presentation here has spelled out, in considerable detail, the kinds of issues that researchers constructing a new measure must consider. There is no simple formula, but the integrated conception of construct validity and the various validation procedures summarized in Table 18.4 provide a blueprint for the kinds of evidence to be gathered and procedures to be followed.

Questionnaire construction, like measurement more generally, involves theory-building and thus requires an iterative process. It begins with (a) generating hypotheses; (b) building a model and plausible alternatives; (c) generating items using construct definitions, generalizability facets, and content validation procedures as guides (for information about item and response formats, see Visser, Krosnick, Lavrakas, & Kim, Chapter 16 in this volume); (d)

gathering and analyzing data; (e) confirming and disconfirming the initial models; and (f) generating alternative hypotheses leading to improved models, additional and more content-valid items, more data gathering, and so on. The cycle continues, until a working model has been established that is “good enough” – one that the investigator can live with, at least for a while, given the constraints and limits of real-life research.

Cultural and Translation Issues in Questionnaire Construction. After years of relative neglect, interest in cross-cultural research has been growing in the 1990s (Van de Vijver & Leung, 1997). There are both theoretical and practical reasons for examining psychological measures in cultures other than the United States. First, cross-cultural studies are needed to test the generalizability of our psychological theories and models. Second, given the increasing multiculturalism in the United States, cultural research is necessary to understand the psychological reality of cultural and ethnic minorities.

Methodological considerations are very important in cross-cultural research. Consider a researcher who wants to test a theory that two culture groups differ on a measure or that they show different correlates with a measure. First, the researcher must demonstrate that the same characteristic has been measured in the same way across the two groups. The most common research strategy has been to translate an original U.S.-developed measure to assess the construct of interest in a new culture. This *imposed-etic* strategy (Berry, 1980) is economical and efficient when we want to examine how a particular measure generalizes to other cultures. However, when we want to identify culture-specific aspects of a construct, the imposed-etic approach has serious limitations; using translated measures simply assumes that the construct is universal, thus ignoring meanings and indicators of the construct that are potentially culture-specific (Church & Katigbak, 1988).

The question whether imposed-etic measures overlook important domains of the local culture is at the core of the long-standing *emic-etic* debate (Berry, Poortinga, Segall, & Dasen, 1992), which contrasts the supposedly interculturally comparable, universal (etic) aspects of a construct with its culture-specific, indigenous (emic) aspects (Berry, 1980). On the one hand, an imposed-etic strategy is useful in that it makes cross-cultural comparisons feasible (i.e., statements about the similarity of two cultures require dimensional equivalence), yet its use may distort the meaning of the construct. On the other hand, a fully emic strategy is well suited to identify culture-specific aspects of a construct (i.e., it is ecologically valid), but it renders comparisons across cultures

virtually impossible (Berry, 1980). Note that the emic-etic distinction is not “either-or” but a matter of degree. Overlap between measures taken in different cultures is not simply present or absent, but rather varies in strength and breadth (Berry et al., 1992).

The current view is that emic and etic approaches render two distinct (though related) types of information. Thus, the two approaches need to be combined to provide a complete picture of cultural specificity and overlap (Benet-Martínez & Waller, 1997; Church & Katigbak, 1988; Yang & Bond, 1990). The use of a combined emic-etic approach requires the researcher (a) to identify the emic (indigenous) elements of the construct (through focus groups, interviews, or content analyses of popular media), and develop and administer measures that adequately tap these constructs; (b) to administer translated measures tapping imposed-etic constructs; and (c) to assess the specificity and overlap between imported and indigenous measures. By comparing the information yielded by imposed-etic and emic measures, the researcher can assess how well imported and indigenous constructs correspond and identify indigenous elements not represented by the imported (translated) instrument (for an illustration of this procedure, see Benet-Martínez & Waller, 1997).

An indispensable requirement for valid cross-cultural comparisons is conceptual equivalence, that is, symmetry in the meaning of different-language versions of a measure (see Van de Vijver & Leung, 1997 for a discussion of construct, measurement, and scalar equivalence in cross-cultural comparisons). One way to foster conceptual equivalence is to use the back-translation procedure (Brislin, 1980). One fluent bilingual (ideally an expert on the construct of interest) translates the instrument from the original language into the language of interest. A second bilingual expert independently translates these materials back into the original language. The combination of (a) comparing the back-translated version to the original, (b) discussions between translators, and (c) back-and-forth translations should lead to a final set of translated items that are symmetrically translatable to the original language counterparts.

Following careful back-translation procedures, construct validation procedures must be used to check the success of the translation. In comparisons of two monolingual samples (e.g., one Spanish speaking, the other English speaking), discrepancies in item and scale statistics indicate lack of equivalence but fail to reveal its source – it might be because of poor translations, but sample and culture differences could also play a role. Thus, ideally, the two language versions are compared across samples of both monolinguals and bilinguals (see

Benet-Martínez & John, 1998 for an illustration of how a bilingual design can be used to disentangle these confounds). Most recently, IRT (see our earlier discussion) has become an effective and popular tool for examining cross-cultural and cross-linguistic measurement invariance (e.g., Ellis, Becker, & Kimmel, 1993). If sample sizes are large enough, measurement invariance across languages can also be tested with CFA (Benet-Martínez & John, 1998). Together, CFA and IRT hold much promise to help resolve these special measurement problems in cross-cultural research (see Reise, Widaman, & Pugh, 1993).

Conclusions and Recommendations

In this chapter we have tried to strike a balance between description and prescription, between “what is” and “what should be” the practice of measurement in social-personality research. We reviewed the traditional reliability coefficients but urged the reader to think about facets of generalizability, such as time, items, and observers, and to explicitly adopt a generalizability framework. We railed against some of our pet peeves, such as the indiscriminate use of alpha, pointing out its limitations and arguing for more complex interpretations of this ubiquitous index. We discussed a unified conception of construct validity, suggesting that systematic construct validation efforts are needed to develop a theoretical understanding of our methods; this goal is worth a sustained program of research rather than a few isolated criterion correlations sprinkled throughout the literature. We noted the voracious appetite our field has for “fast data” (the so easily obtained self-reports) and argued for a more diversified diet, calling for multimethod investigations as a rule, rather than the rare exception. We illustrated, briefly, the power of the no-longer new SEM techniques to address measurement problems, calling for their routine use, at least in samples of large size (of which we would like to see more, too).

This chapter has noted shortcomings in the current practice of measurement that one could deplore, as well as practices that ought to be changed. Nonetheless, we are upbeat about the future of measurement in social-personality psychology. In writing this chapter we became particularly persuaded by the simple logic of comparative model-testing: We now see it as the best strategy for evaluating and improving our measurement procedures. We are confident that even though our journals still practice the archaic preoccupation with significance tests, comparative model-testing will catch on, eventually, and so will the powerful tools provided by SEM. Of course, it will

not happen overnight. As Jacob Cohen (1990) concluded from his 40 years of research on methodology, the “inertia” of methodological advance is enormous “but I do not despair...these things take time” (p. 1311).

References

- Altemeyer, R. (1988). *Enemies of freedom: Understanding right-wing authoritarianism*. San Francisco, CA: Jossey-Bass.
- American Psychological Association. (1954). Technical recommendations for psychological tests and diagnosis techniques. *Psychological Bulletin*, 51, 201–238.
- American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Benet-Martínez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, 75, 729–750.
- Benet-Martínez, V., & Waller, N. G. (1997). Further evidence for the cross-cultural generality of the “Big Seven” model: Imported and indigenous Spanish personality constructs. *Journal of Personality*, 65, 569–598.
- Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, 31, 419–456.
- Berry, J. W. (1980). Introduction to methodology. In H. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (Vol. 2, pp. 1–28). Boston: Allyn & Bacon.
- Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (1992). *Cross-cultural psychology: Research and applications*. New York: Cambridge University Press.
- Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, 117, 187–215.
- Block, J., & Robins, R. W. (1993). A longitudinal study of consistency and change in self-esteem from early adolescence to early adulthood. *Child Development*, 64, 909–923.
- Block, J. H., & Block, J. (1980). The role of ego-control and ego-resiliency in

- the organization of behavior. In W. A. Collins (Ed.), *Development of cognition, affect, and social relations: The Minnesota symposia on child psychology* (Vol. 13, pp. 40–101). Hillsdale, NJ: Erlbaum.
- Bollen, K. A. (1984). Multiple indicators: Internal consistency or no necessary relationship? *Quality and Quantity*, 18, 377–385.
- Bollen, K. A., & Long, J. S. (Eds.). (1993). *Testing structural equation models*. Newbury Park, CA: Sage.
- Brislin, R. W. (1980). Translation and content analysis of oral and written materials. In H. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (Vol. 2, pp. 389–444). Boston: Allyn & Bacon.
- Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist*, 39, 214–227.
- Burisch, M. (1986). Methods of personality inventory development – A comparative analysis. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaire* (pp. 109–120). Berlin, Germany: Springer-Verlag.
- Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42, 116–131.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Cattell, R. B. (1957). *Personality and motivation structure and measurement*. New York: World Book.
- Cattell, R. B. (1972). *Personality and mood by questionnaire*. San Francisco, CA: Jossey-Bass.
- Chaplin, W. F., John, O. P., & Goldberg, L. R. (1988). Conceptions of states and traits: Dimensional attributes with ideals as prototypes. *Journal of Personality and Social Psychology*, 54, 541–557.
- Church, A. T., & Katigbak, M. S. (1988). The emic strategy in identification and assessment of personality dimensions in a non-western culture. *Journal of Cross-Cultural Psychology*, 19, 140–163.
- Cliff, N. F. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, 3, 186–190.

- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304–1312.
- Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98–104.
- Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R. The Revised NEO Personality Inventory*. Odessa, FL: Psychological Assessment Resources.
- Cronbach, L. J. (1947). Test “reliability”: Its meaning and determination. *Psychometrika*, 12, 1–16.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Cronbach, L. J., & Gleser, G. C. (1957). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Cronbach, L. N., Rajaratnam, N., & Gleser, G. C. (1963). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, 25, 291–312.
- Dawes, R. M. (1994). Psychological measurement. *Psychological Review*, 101, 278–281.
- Dawes, R. M., & Smith, T. L. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 1, pp. 509–566). New York: Random House.
- Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Ft. Worth, TX: Harcourt Brace Jovanovich.
- Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory (TPI).

- Journal of Cross-Cultural Psychology*, 24, 133–148.
- Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 8, 341–349.
- Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist*, 35, 790–806.
- Feldt, L., Woodruff, D., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, 11, 93–103.
- Fishbein, M., & Ajzen, I. (1974). Attitudes towards objects as predictors of single and multiple behavioral criteria. *Psychological Review*, 81, 59–74.
- Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7, 286–299.
- Goldberg, L. R. (1971). A historical survey of personality scales and inventories. In P. McReynolds (Ed.), *Advances in psychological assessment* (Vol. 2, pp. 293–336). Palo Alto, CA: Science and Behavior Books.
- Goldberg, L. R. (1991). Clinical versus statistical prediction. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul E. Meehl* (pp. 173–184). Minneapolis: University of Minnesota Press.
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48, 26–34.
- Goldberg, L. R., & Kilkowski, J. M. (1984). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology*, 48, 82–98.
- Gough, H. G. (1987). *The California Psychological Inventory administrator's guide*. Palo Alto, CA: Consulting Psychologists Press.
- Gould, S. J. (1981). *The mismeasure of man*. New York: W.W. Norton.
- Gray-Little, B., Williams, S. L., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, 23, 443–451.
- Gross, J. J., & John, O. P. (1997). Revealing feelings: Facets of emotional

expressivity in self-reports, peer ratings, and behavior. *Journal of Personality and Social Psychology*, 72, 435–448.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin*, 69, 192–203.

Hampson, S. E. (1998). When is an inconsistency not an inconsistency? Trait reconciliation in personality description and impression formation. *Journal of Personality and Social Psychology*, 74, 102–117.

Harris, G. T., & Rice, M. E. (1996). The science in phallometric measurement of male sexual interest. *Current Directions in Psychological Science*, 5, 156–160.

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory* (rev. ed.). Minneapolis: University of Minnesota Press.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9, 139–164.

Himmelfarb, S. (1993). The measurement of attitudes. In A. H. Eagly & S. Chaiken (Eds.), *The psychology of attitudes* (pp. 23–87). Ft. Worth, TX: Harcourt Brace Jovanovich.

Hull, J. G., Lehn, D. A., & Tedlie, J. C. (1991). A general approach to testing multi-faceted personality constructs. *Journal of Personality and Social Psychology*, 61, 932–945.

Jackson, D. N. (1971). The dynamics of structured personality tests: 1971. *Psychological Review*, 78, 229–248.

Jackson, D. N. (1984). *Personality Research Form manual*. Port Huron, MI: Research Psychologists Press.

John, O. P. (1990). The Big Five factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66–100). New York: Guilford Press.

John, O. P., Cheek, J. M., & Klohnen, E. C. (1996). On the nature of self-

- monitoring: Construct explication via Q-sort ratings. *Journal of Personality and Social Psychology*, 71, 763–776.
- John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The “Big Five” Inventory: Versions 4a and 54* [Technical Report]. Institute of Personality and Social Research, University of California, Berkeley.
- John, O. P., Hampson, S. E., & Goldberg, L. R. (1991). The basic level in personality-trait hierarchies: Studies of trait use and accessibility in different contexts. *Journal of Personality and Social Psychology*, 60, 348–361.
- John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality*, 61, 521–551.
- John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology*, 66, 206–219.
- John, O. P., & Srivastava, S. (1999). The Big Five taxonomy: History, measurement, and theoretical perspective. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). New York: Guilford Press.
- Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: User's guide*. Chicago: National Educational Resources.
- Judd, C. M., Jessor, R., & Donovan, J. E. (1986). Structural equation models and personality research. *Journal of Personality*, 54, 149–198.
- Judd, C. M., & McClelland, G. H. (1998). Measurement. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 2, pp. 180–232). Boston: McGraw-Hill.
- Judd, C. M., McClelland, G. H., & Culhane, S. E. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. *Annual Review of Psychology*, 46, 433–465.
- Kagan, J., & Snidman, N. (1991). Infant predictors of inhibited and uninhibited profiles. *Psychological Science*, 2, 40–44.
- Kelderman, H., & Rijkes, C. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, 59, 149–176.

- Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology*, 12, 247–252.
- Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford Press.
- Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112, 165–172.
- King, J. E., & Figueredo, A. J. (1997). The five-factor model plus dominance in chimpanzee personality. *Journal of Research in Personality*, 31, 257–271.
- Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis* (3rd ed.). Mahwah, NJ: Erlbaum.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635–694.
- Loevinger, J. (Ed.). (1998a). *Technical foundations for measuring ego development*. Mahwah, NJ: Erlbaum.
- Loevinger, J. (1998b). Completing a life sentence. In P. M. Westenberg, L. Cohn, & A. Blasi (Eds.), *Personality development: Theoretical, empirical, and clinical investigations of Loevinger's conception of ego development* (pp. 347–354). Mahwah, NJ: Erlbaum.
- Lord, F. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement*, 21, 239–243.
- Lord, F., & Novick, M. R. (1968). *Statistical theories of mental tests*. New York: Addison-Wesley.
- Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. *Journal of Personality and Social Psychology*, 67, 808–817.
- Marsh, H. W., Byrne, B. M., & Shavelson, R. J. (1992). A multidimensional, hierarchical self-concept. In T. M. Brinthaup & R. P. Lipka (Eds.), *The self: Definitional and methodological issues* (pp. 44–95). Albany: State University of New York Press.
- McArdle, J. J. (1996). Current directions in structural factor analysis. *Current Directions in Psychological Science*, 5, 11–18.

- McGraw, K.O., & Wong S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30–46.
- Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, 1, 293–299.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
- Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50, 741–749.
- Mischel, W. (1990). Personality dispositions revisited and revised: A view after three decades. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 111–134). New York: Guilford Press.
- Moffitt, T. E. (1993). Adolescence-limited and life-course persistent antisocial behavior: A developmental taxonomy. *Psychological Review*, 80, 252–283.
- Myers, I. B., & McCaulley, M. H. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologists Press.
- Neuberg, S. L., Judice, T. N., & West, S. G. (1997). What the need for closure scale measures and what it does not: Toward differentiating among related epistemic motives. *Journal of Personality and Social Psychology*, 72, 1396–1412.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Ozer, D. J. (1989). Construct validity in personality assessment. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 224–234). New York: Springer-Verlag.
- Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, 66, 1025–1060.
- Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.
- Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag.

- Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, 114, 552–566.
- Rezmovic, E. L., & Rezmovic, V. (1981). A confirmatory factor analysis approach to construct validation. *Educational and Psychological Measurement*, 41, 61–72.
- Robins, R. W., & Hendin, H. M. (1999). *A single item measure of self-esteem: Evidence for its reliability and validity*. Manuscript submitted for publication.
- Robins, R. W., & John, O. P. (1997). The quest for self-insight: Theory and research on the accuracy of self-perception. In H. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 649–679). New York: Academic Press.
- Rosenberg, M. (1979). *Conceiving the self*. New York: Basic Books.
- Schmidt, F. L., Hunter, J. E., Pearlman, K., & Hirsch, H. R. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology*, 38, 697–798.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8, 350–353.
- Schwarzer, R. (1986). Evaluation of convergent and discriminant validity by use of structural equations. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaire* (pp. 192–213). Berlin, Germany: Springer-Verlag.
- Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 44, 922–932.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.
- Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment*, 7, 300–308.
- Snyder, M. (1987). *Public appearances, private realities: The psychology of self-monitoring*. New York: Freeman.
- Sternberg, R. J. (1985). Human intelligence: The model is the message. *Science*,

230, 1111–1118.

- Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.
- Taylor, S. E., & Brown, J. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, 103, 193–210.
- Tellegen, A. (1985). Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In A. H. Tuma & J. Maser (Eds.), *Anxiety and the anxiety disorders* (pp. 681–706). Hillsdale, NJ: Erlbaum.
- Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
- Tinsley, H. E., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology*, 34, 414–424.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
- Van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.
- Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the Absorption scale. *Journal of Personality and Social Psychology*, 57, 1051–1058.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. B. (1981). *Nonreactive measures in the social sciences* (2nd ed.). Boston: Houghton-Mifflin.
- West, S. G., & Finch, J. F. (1997). Measurement and analysis issues in the investigation of personality structure. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 143–164). Dallas, TX: Academic Press.
- Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Menlo Park, CA: Addison-Wesley.
- Yang, K., & Bond, M. H. (1990). Exploring implicit personality theories with

indigenous and imported constructs: The Chinese case. *Journal of Personality and Social Psychology*, 58, 1087–1095.

* Preparation of the earlier edition of this chapter was supported, in part, by Grant MH49255 and 43948 from the National Institute of Mental Health and a sabbatical award from the University of California, Berkeley, to the first author. We are indebted to Harry Reis and Chick Judd for their enormous editorial efforts on behalf of this chapter and to Lewis R. Goldberg and Richard W. Robins for their helpful comments on an earlier draft. Correspondence may be addressed to Oliver P. John, Department of Psychology, University of California, Berkeley, CA 94720-1650 (e-mail: o_johnx5@berkeley.edu).

¹ As noted in [Table 18.1](#), both Pearson and intraclass correlations can be used to index stability. Pearson correlations only reflect changes in the relative standing of participants from one time to the other, which is typically the prime concern in research on individual differences. When changes in mean levels or variances are of interest, too, then the intraclass correlation is the appropriate index.

² Consistent with Schmitt ([1996](#)), the examples are presented in correlational terms (rather than in covariance terms) simply for ease of interpretation and convenience. Alphas are in fact standardized alphas (i.e., after standard scoring all variables).

³ Cortina's ([1993](#)) index should not be confused with the standard error of alpha, which can be computed under certain distributional assumptions (cf. Feldt, Woodruff, & Salih, [1987](#)).

⁴ There is an important exception where this internal-consistency conception does not apply. In most social-personality measurement, the indicators of a construct are seen as effects caused by the construct; individuals endorse particular attitude statements because of underlying individual differences in attitudes. However, as Bollen ([1984](#)) noted, constructs such as socioeconomic status (SES) are different. SES indicators, such as education and income, cause changes in SES, rather than SES causing changes in education or income. In

these cases of “cause indicators,” the indicators are not necessarily correlated and the internal-consistency conception does not apply.

Chapter nineteen Exploring Causal and Noncausal Hypotheses in Nonexperimental Data

Leandre R. Fabrigar and Duane T. Wegener

A great deal of psychological inquiry is based on studies that do not use experimental manipulations. In this chapter we review statistical methods used to explore causal (directional) and noncausal research questions in such settings. Because complete coverage of these techniques is not feasible in a single chapter, our goal is to introduce readers to some of the techniques that are commonly used or have the potential to be especially useful for social-personality psychology researchers. For each method, we want to help readers understand the sorts of questions that can be addressed by the method, major issues researchers face in its application, and how it relates to other methods. We also provide references for more detailed discussions of each method.

Why Conduct Nonexperimental Studies?

Many of the contexts in which psychologists use nonexperimental data can be placed into four broad categories – situations in which (a) the question(s) of interest do not involve causal relations, (b) the variables of interest cannot or should not be manipulated, (c) nonexperimental procedures increase the efficiency and expediency of research, and (d) the goal is to determine if a previously demonstrated experimental effect can be generalized to a more naturalistic setting (Brewer & Crano, Chapter 2 in this volume). Research questions that do not involve causal relations often include explorations of the number and nature of distinct psychological dimensions thought to underlie a domain of interest. For instance, classic studies on the Big Five theory of personality explored the number and nature of the fundamental dimensions underlying personality descriptors (see John & Srivastava, 1999). Other “noncausal” research includes studies aimed at developing and validating a measure for a particular construct(s). In these examples, the primary goal of the researcher is to determine how many dimensions underlie a set of variables and to understand the psychological nature of the dimensions.

In other contexts, researchers might have hypotheses that are causal in nature, but might collect nonexperimental data because they are interested in variables that cannot be manipulated. For example, personality traits by their nature cannot be readily manipulated. Likewise, demographic variables such as sex, ethnicity, and age cannot be changed by experimenters. In such cases, researchers measure natural variation of variables and assess the hypothesized relations. In some cases it might be conceptually possible to manipulate variables, but practical or ethical issues preclude it. For example, in principle it would be possible to randomly assign children to be raised in complete isolation or with extensive human contact, but such a manipulation would be extremely difficult to implement and unethical to attempt.

Even in situations in which manipulation is feasible and desirable, nonexperimental aspects of the study are often useful. For example, hypotheses involving complex consequences of a given predictor variable can often be usefully explored using both experimental and nonexperimental approaches. Consider a researcher interested in the role of three possible mediators of the impact of a manipulated independent variable (IV) on a dependent variable (DV). It might be possible to examine each stage of this chain through four separate manipulations (one examining the impact of the IV on the three mediators and three other manipulations, each examining the impact of one of the mediators on the DV). However, it might be more efficient and feasible to use statistical procedures assessing mediation in a study in which the IV is manipulated but the mediators are measured (though at some stage, the manipulation of proposed mediators would provide greater leverage in concluding that the proposed mediators have a causal impact on the DV; Judd, Yzerbyt, & Muller, Chapter 25 in this volume; Spencer, Zanna, & Fong, 2005). Later in the chapter we discuss procedures for testing mediation hypotheses.

“Experimental” Versus “Nonexperimental” Statistical Methods

Researchers have often regarded some statistical procedures as “experimental” (e.g., Analysis of Variance) and other statistical procedures as “nonexperimental” (e.g., correlation, structural equation modeling [SEM]). This distinction is highly problematic in that many statistical methods can be readily applied to both experimental and nonexperimental data.¹ The choice of statistical procedure is more a function of the research question than the experimental

versus nonexperimental nature of the design, and the nature of the conclusions one can make about one's data are often more a function of the research design than the statistical procedure. Thus, it seems more useful to organize analysis procedures according to the types of research questions they can address. Some statistical methods address only noncausal hypotheses involving relations among the constructs of interest (e.g., exploratory factor analysis, multidimensional scaling). Other procedures (e.g., multiple regression, hierarchical linear modeling) generally address directional (causal) hypotheses. Many procedures can be used to assess directional or nondirectional hypotheses (e.g., SEM). It is important to note that directional hypotheses might be tested with data collected via experimental or nonexperimental means. Directional models applied to nonexperimental data might sometimes provide a less than compelling case for attributing cause to the “causal” factor in the model. However, causes are modeled in those analyses. As we discuss later, when threats to causal inference are minimal (e.g., existence of alternative explanations or alternative mathematically equivalent models; MacCallum, Wegener, Uchino, & Fabrigar, 1993), some directional analyses can provide a reasonably strong basis for making causal inferences.

Analyses Addressing Noncausal Hypotheses

Each discussion of a “noncausal” procedure begins with an overview of the goals of the statistical technique. Then, we discuss practical issues that influence the implementation of the analysis and design features of the study relevant to that statistical procedure.

Exploratory Factor Analysis

Beyond simple correlations, perhaps the most widely used noncausal analysis in social-personality psychology is exploratory factor analysis (EFA).² The objective of EFA is to identify the number and nature of common factors underlying a set of measured variables. EFA is most commonly used when a researcher is attempting to develop a theory about the latent variables (constructs) in a conceptual domain and/or to construct a scale to assess a set of constructs. In theory development, a researcher might be interested in a particular class of variables and wish to identify the basic constructs underlying the variables. For instance, in research on attitude structure, researchers could have investigated the tripartite theory (e.g., Rosenberg & Hovland, 1960) using EFA to investigate the existence of three latent types of evaluative responding –

affect, cognition, and behavior (see Fabrigar, Wegener, MacCallum, & Strahan, 1999). In scale development, a researcher might wish to identify the dimensionality of a battery of items and the relation(s) of these items to the dimensions (e.g., see development of a scale measuring affective vs. cognitive bases of attitudes in Crites, Fabrigar, & Petty, 1994). EFA is used when the researcher does not have strong a priori predictions or when the researcher has incompletely formulated expectations regarding the number of constructs or their relations to the measured variables (see Fabrigar & Wegener, 2012).

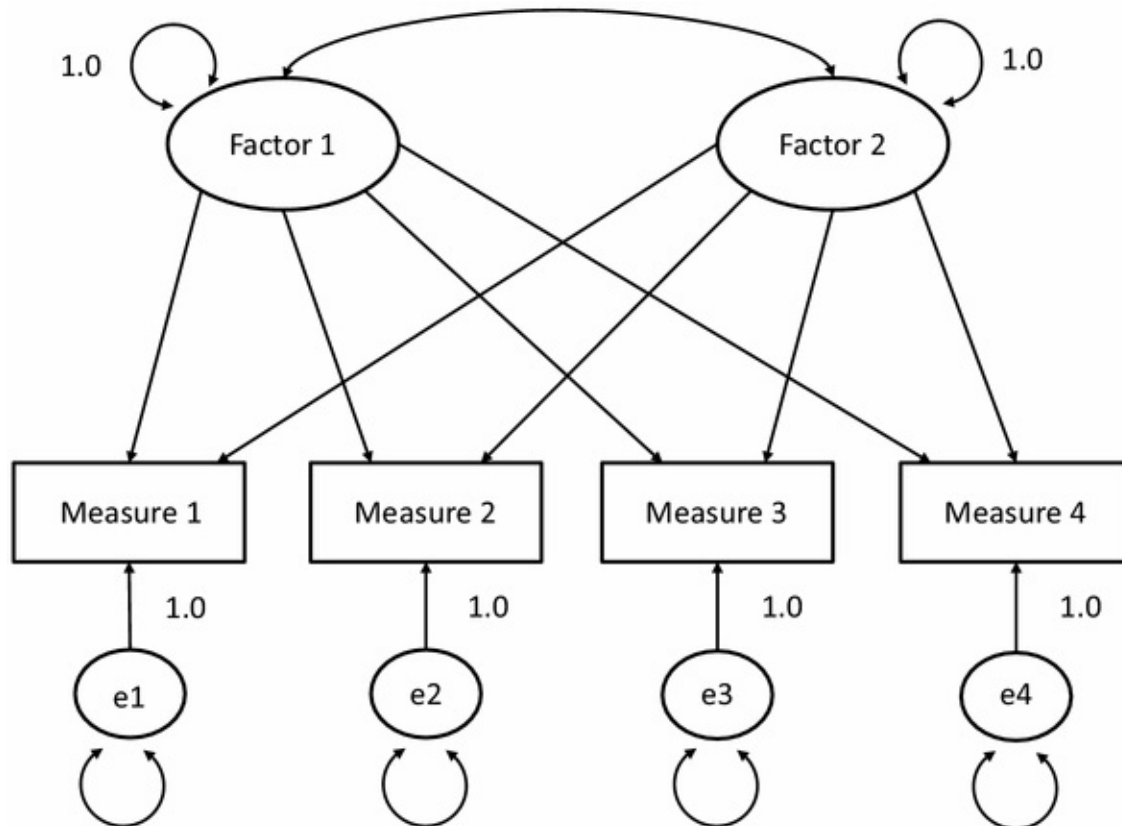


Figure 19.1. Path diagram of an EFA of four measures extracting two common factors. *Note:* Common factors (i.e., latent variables) are represented by ellipses and measures are represented as rectangles. Error terms (unique variances) are displayed as latent variables using circles or ellipses. Directional relations between variables are represented as single-headed arrows and nondirectional relations as doubleheaded arrows. Variances of latent variables are represented by doubleheaded arrows from the variable to itself. Arrows without numerical values are free parameters, and numerical values of the arrows, or paths, are values of fixed parameters specified by the researcher.

Statistical Issues in Conducting EFA.

EFA begins with the matrix of correlations (or covariances) among measured variables. A model with a particular number of factors is specified and fit to the matrix of correlations. This process involves estimating model parameters that minimize the discrepancy between the predicted (model) and observed correlation matrices. These estimates include loadings of the measured variables on the factors (i.e., the strength of the relations between factors and measures), communalities (i.e., the amount of variance in each measure accounted for by the factors), and correlations among factors (in oblique rotations). Based on these parameter estimates, especially the factor loadings, the researcher interprets the nature of the factors according to the measured variables most strongly associated with each factor. [Figure 19.1](#) shows a factor analytic model in which two common factors are hypothesized to underlie the correlations among four measured variables. As depicted, in EFA, each factor can potentially influence all of the measured variables.³ EFA solutions are typically rotated in order to aid interpretation and simplify representation. Therefore, loadings (relations between the factors and the measured variables) for a given factor will typically be high for some subset of the measured variables. Perhaps more than any other commonly used statistical method in psychology, EFA requires a researcher to make a number of important decisions.

Factor extraction. Factor extraction involves finding values for the model parameters (i.e., factor loadings and communalities) that produce correlations predicted by the model that comes as close as possible to the observed set of correlations. The most widely available factor extraction procedures are noniterated principal axis (NIPA), iterated principal axis (IPA), and maximum likelihood (ML) factor analysis. All three procedures are based on the common factor model. Thus, under reasonably good conditions (e.g., strong common factors, proper model specification, and data that do not violate assumptions), the procedures produce very similar results (e.g., see Briggs & MacCallum, [2003](#); Widaman, [1993](#)). However, under less optimal conditions, some differences can be observed.

The strengths of NIPA and IPA are that they do not assume multivariate normality of measured variables and seldom encounter parameter estimation problems. The strengths of ML are that it permits calculation of model fit indices as well as computation of standard errors, confidence intervals, and significance tests for model parameters.⁴ However, ML also has some drawbacks. It assumes multivariate normality of measured variables, though it is relatively robust to violations of this assumption. When moderate amounts of model error and/or

sampling error are present, ML is also less effective than IPA in recovering weak common factors (Briggs & MacCallum, 2003). Because of the additional information provided by ML, we consider ML preferable to IPA and NIPA in most contexts. However, it is always prudent to examine the IPA and ML solutions to confirm that they produce comparable results. Substantial differences can highlight the need to further consider the model being fit and the properties of the data.

Selecting the number of factors. Determining the number of common factors to specify in an EFA model has long been recognized as a great challenge. The appropriate number of common factors is a reflection of both statistical utility (i.e., the number of factors that effectively account for the correlations among measured variables) and conceptual utility (i.e., factors that substantially simplify the data and can be readily interpreted). Because no procedure can fully address both of these concerns, this decision is best addressed by considering the configuration of evidence presented by several of the better performing procedures.

An extensive literature has developed regarding procedures for determining the optimal number of factors. The most widely used (and most flawed) of these procedures is the eigenvalues-greater-than-one rule (the *Kaiser criterion*). This procedure involves computing the eigenvalues from the unreduced (or reduced) correlation matrix (i.e., the variance in the measured variables accounted for by each principal component or common factor) and then examining the number of eigenvalues greater than one. The rule had an underlying logic in its application to eigenvalues from the unreduced correlation matrix (i.e., the correlation matrix among the measured variables) used in PCA. However, it was never intended for use with eigenvalues from the reduced correlation matrix (i.e., the correlation matrix with initial estimates of the communalities of the measured variables in the diagonal), which is the matrix used in EFA. As with any mechanical rule, the Kaiser criterion can lead to arbitrary decisions such as a factor with an eigenvalue of 1.01 being retained but a factor with an eigenvalue of .99 not being retained, even though the difference in variance accounted for by the factors is trivial. Studies have suggested the procedure performs poorly in both PCA and EFA (e.g., Hakstian, Rogers, & Cattell, 1982; Tucker, Koopman, & Linn, 1969).

Another widely used method of determining the number of factors is the *scree test*. This procedure involves plotting the eigenvalues from the unreduced or reduced correlation matrix in descending order. The graph is examined to

determine the number of eigenvalues that precedes the last major drop. This number of factors is specified in the model. When conducting an EFA, it is more sensible to plot the reduced matrix eigenvalues as these are the eigenvalues that more directly correspond to the extracted common factors. Unfortunately, standard statistical programs do not always plot the eigenvalues that correspond to the analysis being conducted (Fabrigar & Wegener, 2012). Although the scree test has been criticized for its subjectivity (e.g., there is no clear definition of “major drop” in a scree plot), studies have suggested that when strong common factors are present in the data, the procedure works reasonably well (Hakstian et al., 1982; Tucker et al., 1969).

Parallel analysis involves calculating the eigenvalues that would be expected from a set of random data with the same number of measured variables and same sample size as the real data set. The eigenvalues from random data are compared with their corresponding eigenvalues from the real data (i.e., the first eigenvalue from real data set is compared with the first eigenvalue from random data, the second eigenvalue from real data is compared with the second eigenvalue from random data, and so on). The appropriate number of common factors is the number of eigenvalues from the real data that are larger than their corresponding eigenvalues from random data. As with the scree test, a researcher can base a parallel analysis on eigenvalues from the reduced correlation matrix or eigenvalues from the sample correlation matrix. In the context of EFA, parallel analyses procedures examining the reduced correlation matrix seem most sensible. Because the parallel analysis criterion is that a factor must simply outperform random data, it is a comparatively lenient standard for factor retention. Thus, this procedure might be viewed as establishing the upper boundary of the number of common factors that should be considered (e.g., see Buja & Eyuboglu, 1992). Nonetheless, parallel analysis procedures perform reasonably well in simulated data sets with strong common factors and in which no minor common factors were present (e.g., Humphreys & Montanelli, 1975).

The three methods described thus far are traditionally used with principal axis factor extraction. More recently, with ML extraction, indices of model fit have been used to determine the optimal number of common factors. The model fit approach involves computing goodness of fit for a series of models with differing numbers of factors – beginning with zero and increasing by one factor until some maximally interesting number. The appropriate number of factors is that number for which (a) the model fits the data well in absolute terms, (b) the model fits substantially better than a model with one fewer factor, and (c) the model fits similarly to a model with one more factor. When using ML extraction,

any of the model fit indices used for confirmatory factor analyses and SEM can be used for EFA (Browne & Cudeck, 1992; see also the later description of fit indices in CFA).

It should be recognized that even the better-performing procedures are not infallible. Each procedure approaches the number-of-factors question in a different manner. Thus, one procedure should never be used in isolation. Additionally, researchers should always consider the interpretability of the solution when determining the number of factors and, when possible, the stability of different solutions across data sets (see Fabrigar & Wegener, 2012).

Factor rotation. For any solution with two or more factors, there exists an infinite number of alternative orientations of the factors in multidimensional space that account equally well for the correlations among measured variables. Therefore, researchers must select a single orientation from among the equally fitting solutions. The criterion most commonly used to guide this decision is *simple structure*. Simple structure refers to solutions in which each factor is defined by a subset of measured variables with large loadings relative to the other measured variables (i.e., high within-factor variability in loadings) and in which each measured variable loads highly on only a subset of the common factors (i.e., low factorial complexity in defining variables). Thus, factors are usually rotated to the solution with the best simple structure.

Numerous procedures for rotating solutions have been proposed. The most fundamental distinction is between orthogonal and oblique rotations. Orthogonal rotations constrain factors to be uncorrelated. Of the orthogonal rotations, varimax is one of the best performing and most widely used. In contrast, oblique rotations permit correlations among factors. A common misconception is that oblique rotations require correlated factors. This is not true. If the best simple structure involves orthogonal factors, successful oblique rotations will estimate interfactor correlations that are close to zero and will produce factor loadings similar to successful orthogonal rotations. However, when the best simple structure includes correlated factors, oblique rotations will produce a solution with correlated factors and generally produce better simple structure than an orthogonal rotation (see Fabrigar & Wegener, 2012 for a graphic illustration).⁵ Thus, oblique rotation is generally a more sensible approach than orthogonal rotation. The three most commonly used oblique rotations are direct quartimin (sometimes called direct oblimin), promax, and Harris-Kaiser orthoblique.

When interpreting oblique rotations, it is useful to keep in mind some differences between oblique and orthogonal rotated solutions. First, in

orthogonal solutions, factor loadings can be interpreted as correlations between the common factors and the measured variables. Thus, they are bounded by -1.00 and 1.00 . In oblique rotations, factor loadings are comparable to standardized partial regression coefficients and thus are not bounded by -1.00 and 1.00 (although they rarely go much beyond these values). A second difference is that an orthogonal rotation will produce a single rotated factor loading matrix. However, it is customary for programs using oblique rotations to report three matrices: the pattern matrix, the structure matrix, and the factor correlation matrix. It is the pattern matrix that corresponds to the rotated factor loading matrix and thus should be the basis of interpretation for the factor loadings (Fabrigar & Wegener, 2012). The structure matrix represents the zero-order correlations between the items and the common factors. Because this matrix does not control for spurious effects of factors on measured variables, in cases where factors are substantially correlated with one another, this matrix will typically show poorer simple structure. The factor correlation matrix is the matrix of correlations among the common factors and can be interpreted as any matrix of correlations, though these are correlations among latent variables and thus reflect associations controlling for attenuation attributable to measurement error. **Summary.** Each decision can influence the results of EFA. This makes it especially important that researchers make their decisions known to readers. Because of the marked differences that can occur when different combinations of procedures are used (Fabrigar et al., 1999), we believe it is essential for authors (with the oversight of editors) to clearly state whether EFA or PCA was used. The precise factor extraction method, means of determining the number of factors, and factor rotation procedure should also be clearly specified. For most situations, an EFA with ML factor extraction, multiple factor number criteria (e.g., scree test, parallel analysis, and model fit), and oblique rotation would be the most sensible approach.

Study Design Issues in EFA.

No statistical procedure can overcome poor design choices. There are two major issues that should be taken into account when designing studies to be analyzed using EFA: selection of measured variables and selection of sample.

Selection of measured variables. Common and unique factors that emerge in any EFA are a direct result of decisions concerning which measured variables to include, as illustrated in Figure 19.2. Four measures are influenced by two common factors and four unique factors. Measures 1 and 2 are primarily

influenced by common factor 1 (the dark solid paths in the figure; the light dotted lines are paths included in the model that showed little or no relation). Measure 3 is influenced by both common factors, and measure 4 is influenced by common factor 2. For example, if common factor 1 was verbal intelligence and common factor 2 was quantitative intelligence, items 1, 2, 3, and 4 might be a reading recall test, vocabulary test, math word problem test, and timed test of multiplication facts. If a researcher had chosen not to include the multiplication test (measured variable 4), a substantially different pattern of results would have emerged. Only common factor 1 (verbal intelligence) would have been obtained, but without the contrast with the other factor, the researcher might not have labeled the single factor as verbal intelligence per se. Common factors are defined as latent variables that influence more than one measured variable. If the multiplication test were excluded, quantitative intelligence would only influence a single measured variable and would have become part of the unique factor (latent variable “e3”) influencing the word problem test measure.

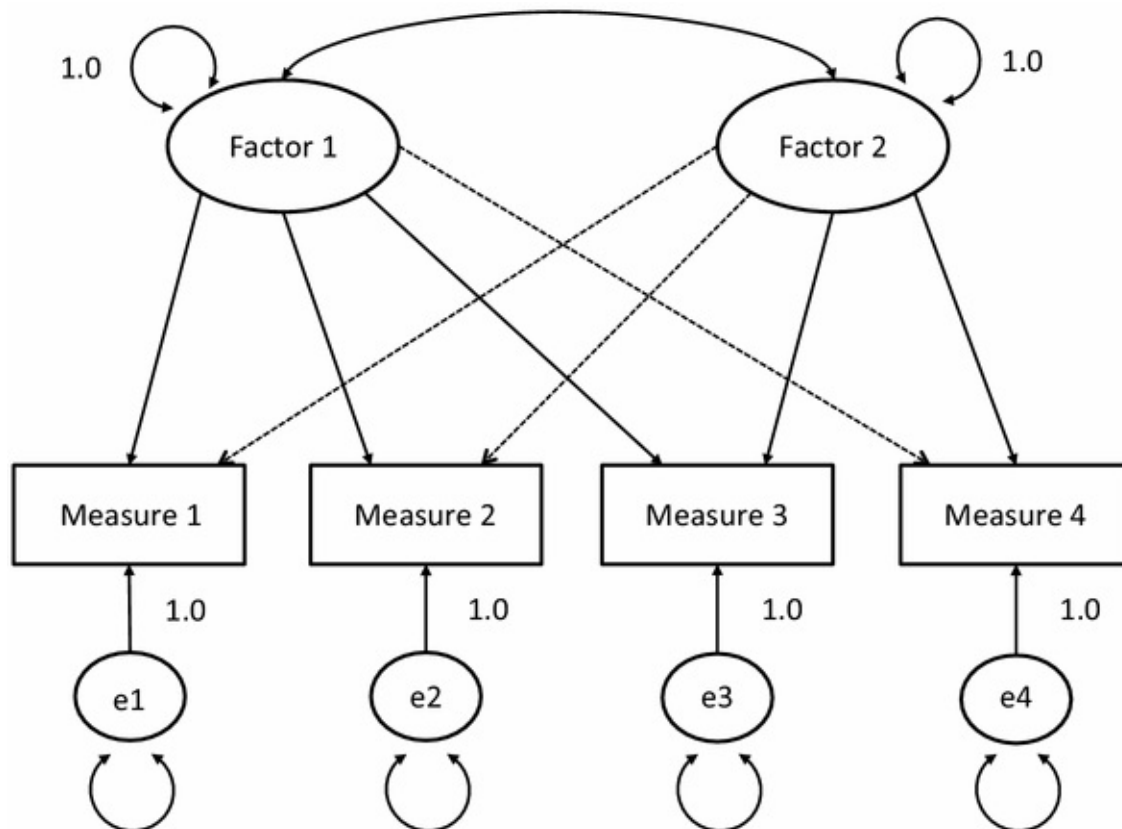


Figure 19.2. Path diagram of an EFA illustrating the impact of sampling of measures from the domain of interest.

Thus, when conducting an EFA, the researcher should carefully define the

area of inquiry, systematically consider the extent to which each measured variable satisfies the conceptual requirements of the area of inquiry, and be sure that the set of measured variables adequately samples the domain. The broader the domain, the greater the number of measured variables required. EFA procedures also perform better when factors are *overdetermined* (i.e., when each factor has multiple measured variables strongly influenced by that factor). At least three to five measured variables should be included to reflect each common factor, although more is generally desirable (MacCallum, Widaman, Zhang, & Hong, 1999). Thus, researchers should include more than five measured variables to represent each possible factor (in the event that some measured variables do not load on their expected factor). Additionally, EFA procedures function better when the communalities of measured variables are high. One cause of low communalities is measures with substantial random error. Thus, sound measurement practices should be used when designing or selecting measured variables (Fabrigar & Wegener, 2012).

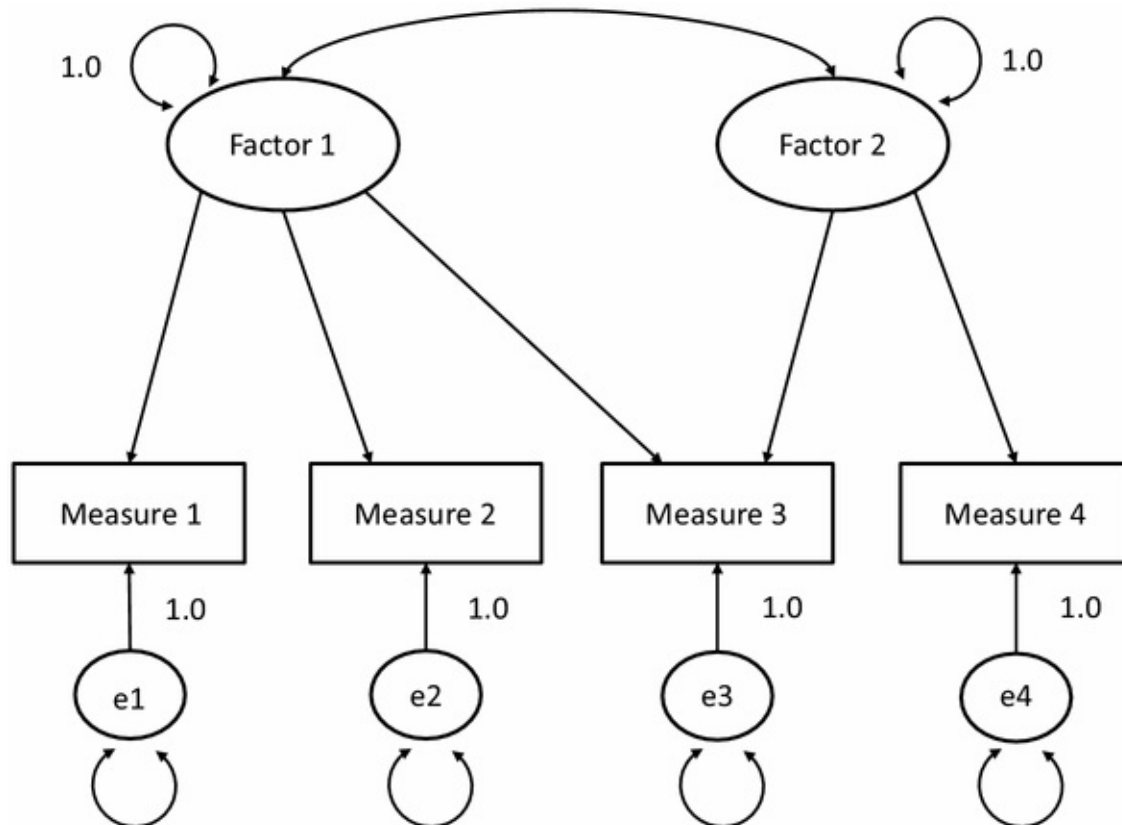


Figure 19.3. Path diagram of a CFA of four measures specifying two correlated factors.

Selection of research participants. Psychologists often select samples for

convenience (e.g., student samples) rather than representativeness of the population as a whole. In many cases, this would not pose problems. However, if the sample is substantially more homogeneous than the population on one or more common factors, variance is reduced and can attenuate the correlations among measured variables, yielding falsely low estimates of the factor loadings and interfactor correlations (Gorsuch, 1983; Tucker & MacCallum, 1997).

A second issue related to sampling is the number of participants needed. Textbooks on factor analysis and multivariate statistics routinely report rules of thumb, usually based on a ratio of participants to measured variables. These guidelines never had strong theoretical or empirical foundations, but were based largely on intuition. Subsequent research has revealed them to be flawed (MacCallum et al., 1999). The sample size required to obtain accurate results depends not on the number of measured variables being analyzed, but instead on the communalities of the variables and their ratio to the number of common factors. Fabrigar *et al.* (1999) suggested that, under relatively optimal conditions (communalities of .70 or greater and at least 3 to 5 measured variables loading on each factor), a sample of 100 can be adequate. Indeed, under extremely optimal conditions, reasonable results can sometimes even occur with smaller samples (less than 50; Preacher & MacCallum, 2002). Under moderately good conditions (communalities of .40 to .70 and at least 3 measured variables loading on each factor), a sample of at least 200 should suffice. Under poor conditions (communalities lower than .40 and some factors with only 2 measured variables loading on them), samples of at least 400 might be necessary, though it may be that even very large samples are inadequate.

Confirmatory Factor Analysis

Another widely used form of nondirectional analysis is confirmatory factor analysis (CFA). CFA is discussed in some detail elsewhere in this volume (see Widaman & Grimm, Chapter 20 in this volume). Thus, we provide a relatively abbreviated discussion of the topic, with an emphasis on its similarities and differences with EFA.

Comparing CFA and EFA.

CFA is an extension of EFA and is similar in a number of ways. CFA is usually based on the same underlying conceptual model – the common factor model. In both cases, the primary objective is to identify the number and nature of factors underlying a set of measured variables. CFA is sometimes used in the later

stages of scale development and validation, as well as in the exploration of more substantive theoretical questions.

Despite these similarities, CFA differs from EFA in that it requires a priori specification of the number of common factors and the pattern of zero and nonzero loadings of measures on the latent factors. Consider the CFA model in [Figure 19.3](#). In many respects, this model parallels the EFA model in [Figure 19.1](#). However, rather than all measures potentially loading on all factors (rotating to simple structure in EFA creates a pattern of high vs. low loadings, but there are loadings on all factors), this CFA model specifies that factor 1 can only influence three particular measures and factor 2 can only influence two particular measures. Note that the measures influenced by each factor may be nonoverlapping but need not be.

This relatively unrestricted nature of EFA and the more restricted nature of CFA highlights that the methods are typically used in different stages of a research program. When there are no expectations regarding the number of common factors and which measured variables will be influenced by the same common factors, EFA is more advisable. In contrast, when there is a compelling basis to specify a precise number of factors and exactly which measured variables each factor should influence, CFA is generally preferred. Obviously, many situations fall between these two extremes. Sometimes a researcher may have a very general idea regarding how many factors might emerge and some expectations regarding which measured variables will be influenced by these factors. However, the theory and data supporting these expectations might be insufficient to specify with confidence the exact number of factors and to make predictions regarding how every measured variable will be influenced by the factors. Similarly, when generating new measures, the researcher might not be sure that the new items will be influenced only by the intended construct. In such cases, EFA may be more appropriate.

In some cases, it might not be possible to confidently identify a single model as a preferred model, but there might be a basis to precisely specify two or three competing models. As long as there are only a few competing models and the theories motivating the models are developed enough that the competing models can be fully specified, a CFA is usually preferable to EFA. However, with an increasingly large number of competing models, EFA may be more sensible because of the unwieldiness of fitting and comparing a large number of models in CFA. Moreover, a large number of plausible competing models may suggest that substantial ambiguity exists in the area of inquiry and that all potentially

plausible models have yet to be identified.

Regardless of whether EFA or CFA is adopted, it is important to recognize that differences between the approaches are more a function of emphasis than a fundamental difference in goals and underlying assumptions. In fact, many presumed differences between EFA and CFA are more illusory than real. For instance, some researchers have suggested that an advantage of CFA is the ability to formally quantify model fit, compare competing models with respect to their fit, and conduct statistical tests of parameter estimates. However, many of these perceived differences are simply a function of the model fitting procedure. As noted earlier, EFA using ML factor extraction permits calculation of fit indices (Browne & Cudeck, 1992) and the computation of standard errors and significance tests for model parameters (Cudeck & O'Dell, 1994). Therefore, these supposed benefits of CFA do not constitute strong reasons to shift from EFA to CFA, especially if existing understanding of the domain and measures is not sufficiently advanced to directly specify all of the relevant alternative models in CFA.

Nonetheless, the difference in emphasis between the two approaches does provide each approach with certain strengths. Because CFA requires a priori model specification and, therefore, a relatively small number of meaningful models can be considered, it is less likely that the CFA analysis will capitalize on chance characteristics in the data. Because EFA places few restrictions on the pattern of loadings, any single EFA solution might capture chance characteristics of the data. This puts a premium on replication of EFA results across samples (where chance characteristics of data should be unlikely to appear across data sets). However, EFA also has strengths. Whereas a single CFA analysis tests only the specified model (and might miss an alternative model – e.g., a model with certain measures loading on more than one common factor), the optimal solution of many potential patterns can be identified in EFA. In many circumstances, it would make sense to use EFA and CFA within the same program of research. For example, an initial EFA might provide a basis for later specification of CFA models, or a large sample could be split and an EFA conducted on half of the sample, with a CFA conducted on the other half. Alternatively, if a CFA model fits poorly, subsequent EFA analyses might suggest alternative models or reasons why the model had poor fit (e.g., poor measures of certain constructs).

CFA obviously differs from EFA in the use of procedures for determining the number of factors. Because CFA involves a priori specification of the number of

factors, most statistical procedures for determining the number of factors used in EFA (e.g., scree test, parallel analysis) are irrelevant to CFA. The one exception to this is the use of model fit to compare models with differing numbers of factors. Such a comparison might be undertaken in CFA. However, the comparison would typically involve comparison of a more restricted set of models than in EFA. Furthermore, comparison between models in EFA does not require the researcher to specify patterns of factor loadings, whereas such comparisons in CFA do require specification of the pattern of loadings for the models being compared. Thus, the comparisons of models in CFA would involve not just a test of whether the number of factors was appropriate, but also whether the precise pattern of loadings across the models was properly specified. Related to this point, rotations are common in EFA but extremely rare in CFA, because the expected simple structure of factor loadings has already been specified.

Finally, CFA generally permits more flexible and focused hypothesis testing of model parameters. For example, in CFA a researcher can test whether certain factor loadings (or interfactor correlations or communalities) are equivalent to one another. It is also possible to conduct tests of the differences of parameter estimates across groups (see Widaman & Grimm, [Chapter 20](#) in this volume). Although it has been possible to conduct tests of model fit in ML versions of EFA since the 1960s, tests of statistical significance of factor loadings and correlations among factors are rather recent developments (Cudeck & O'Dell, [1994](#)). In contrast to CFA, EFA does not allow for tests of equivalence of parameter estimates within a model.

Statistical Issues in CFA

Model specification.

The first step in conducting a CFA is to specify the model(s) to be tested. That is, the researcher must mathematically define the model by specifying which parameters are free (i.e., parameters with unknown values to be estimated from the data, including loadings of measures on the latent factors, unique variances, and interfactor correlations), fixed (i.e., parameters set to a specific numerical value, most typically at zero), and constrained (i.e., parameters with unknown values that are estimated from the data but must hold a specified mathematical relation to one or more other parameters). Several issues must be considered in this process.

The model must be identified (i.e., it must be possible to compute a unique solution). Unfortunately, there is no practical foolproof method for determining if a model is identified. There are some general procedures (such as the order condition or t-rule), mathematical algorithms, and empirical tests that can help reveal identification problems (e.g., Bollen, 1989; Schumacker & Lomax, 2010). Lack of model identification can generally be avoided if the researcher pays attention to the statistical parsimony of the model (i.e., its number of free parameters). A model should include free parameters only when there is a reasonable theoretical or empirical basis for expecting nonzero values. If this is done and rules of model identification are followed (e.g., specifying scales of measurement for latent variables by fixing factor variances to one or fixing one factor loading for each factor to one), lack of model identification can usually be avoided (see Kenny, Kashy, & Bolger, 1998).

During model specification, a researcher should also consider alternative models. Arguing in support of a particular preferred model is most convincing when the preferred model is shown not only to adequately account for the data but also to do so better than other conceptually plausible alternatives. Finding that a model provides an adequate account of the data is not particularly compelling evidence of the value of the model if other conceptually plausible models might perform just as well (or better). Therefore, it is useful at the model-specification stage to propose and evaluate plausible alternative models.

Alternative models might differ from one another in several ways, such as the number of common factors necessary to account for relations among the measures. Some alternative conceptions might specify different patterns of zero and nonzero loadings without varying the number of factors. In other situations, alternative CFA models might utilize the same number of factors and specify the same pattern of zero and nonzero loadings but differ in the constraints on the parameter estimates. For example, the researcher might wish to test whether a given factor correlates more highly with one factor than with another, or might wish to test whether a given item loads more highly on one factor than on another. To test such hypotheses, one could specify an alternative model in which the two parameters (e.g., interfactor correlations) are constrained to be equal and compare this model with one in which the two parameters are not constrained. If the two parameter values differ substantially, then constraining them to be equal will substantially hurt the ability of the model to account for the data. Such comparisons can take advantage of model-comparison procedures for “nested” models.

Fitting the model to the data. After specifying the model(s) of interest, the researcher must select a procedure by which to fit that model to the data. This process is essentially the same as the factor extraction process in EFA.⁶ ML model fitting has been the dominant procedure used in most applications of CFA (and SEM more generally). As noted earlier, ML assumes multivariate normality of the measured variables. Although ML is robust to moderate violations of this assumption (e.g., skew < 2, kurtosis < 7; see West, Finch, & Curran, 1995), more severe violations can be problematic. In such cases, researchers have several options. One approach is to transform the measured variables to produce more normal distributions (e.g., a power function transformation). One drawback of this approach is numerous transformations exist and no single transformation will be best in all contexts. In some cases, no transformation may produce sufficient improvement to satisfy assumptions of normality. Another response is to use another model fitting procedure that does not assume multivariate normality or is more robust to violations of normality. For example, in recent years, substantial progress has been made in developing “robust methods” of parameter estimation that function better under violations of normality (e.g., Zhong & Yuan, 2011).

Evaluating models. After obtaining parameter estimates, researchers must assess the adequacy of a CFA model in accounting for the correlations among the measured variables. Probably the most widely used basis for assessing model adequacy is overall fit. As noted in our discussion of EFA, model fitting involves determining estimates of model parameters that minimize the discrepancy between the model and the sample correlation (or covariance) matrix (telling a researcher how well the model can account for the observed data). If no parameter values lead to small discrepancies between model and data (i.e., if the model fit is poor), then the model is regarded as implausible. Over the years, many fit indices have been developed.

The most commonly reported index is the likelihood ratio or χ^2 goodness-of-fit test (Bollen, 1989). The null hypothesis for the likelihood ratio is that the model holds exactly in the population. Perhaps the most serious drawbacks to the likelihood ratio test are that the hypothesis of exact fit is unrealistic and that the likelihood ratio is inherently sensitive to sample size, such that large samples lead to rejection of virtually all models (MacCallum, 1990).

Because of these concerns, numerous alternative measures of model fit have been developed. These measures are often referred to as descriptive fit indices because they express fit in terms of the magnitude of discrepancy between the

model and the data rather than as a formal hypothesis test of perfect fit. These descriptive fit indices are usually classified into two broad categories: incremental fit indices and absolute fit indices. A second feature of fit indices is that some indices take into account the parsimony of the model whereas others do not.

Incremental fit indices compare a hypothesized model to the fit of a “null model,” typically in which all measures have variances but no covariances with one another. The “variance only” null model has only been a convention. Other baseline models could be, but rarely are, chosen. Probably the best-known and one of the better-performing incremental fit indices is the Tucker-Lewis index (TLI; NonNormed Fit Index, NNFI). This index has been found to perform reasonably well in many simulation studies, and it takes into account model parsimony. Other commonly used incremental fit indices include the Normed Fit Index (NFI; also called the Bentler-Bonett fit index), which does not adjust for model parsimony, and the Incremental Fit Index (IFI), which does take into account model parsimony.

Other fit indices are “absolute fit indices” that express fit in terms of the absolute discrepancy between the model and the data rather than as a comparison between a target model and a baseline model. One such index that has been widely used, functions reasonably well, and accounts for model parsimony is the Root Mean Square Error of Approximation (RMSEA). Other commonly used absolute fit indices include the Standardized Root Mean Square Residual (SRMR) and the Goodness of Fit Index (GFI), neither of which account for model parsimony. The Adjusted Goodness of Fit Index (AGFI) is another widely used measure of fit which does take into account model parsimony.

Ultimately, decisions regarding model evaluation should not be based on a single fit index in that no index is optimal under all conditions and there are different conceptual ways to approach the concept of model fit. However, indiscriminate use of fit indices is also unwise as some indices have performed poorly in simulation studies (e.g., NFI, GFI, AGFI; Hu & Bentler, 1998). Perhaps the most sensible approach is to select one or two of the better-performing indices from each category of model fit (e.g., TLI/NNFI, IFI, RMSEA, SRMR) and evaluate fit examining the performance of the model across these multiple approaches. In addition, examining the residual matrix (i.e., the matrix of residuals between the correlation matrix predicted by the model and the matrix of observed correlations) can also provide valuable insights regarding model fit (Browne, MacCallum, Kim, Andersen, & Glaser,

2002).

Model fit is only one basis on which a model can be evaluated and compared with competing models. Evaluation of models should also be guided by the extent to which parameter estimates are interpretable, plausible, and replicable. For example, multiple cross-boundary estimates (e.g., negative variance estimates) can undermine model interpretability, regardless of overall fit (alone or in comparison with alternative models). In many circumstances, one might justifiably prefer a model with adequate fit and no questionable parameter estimates even if another model with many cross-boundary estimates fits the data a bit better. One might also regard existing theory and data as factors that influence overall preferences for one model over another. For example, if a model includes one or more parameter estimates that fit well with existing theory and research, but an alternative solution provides estimates that are difficult to reconcile with existing data and theory, this might be enough to prefer the former to the latter, especially if the unexpected values have not been replicated.

Finally, even if two models have roughly equivalent fit, a researcher might prefer the more parsimonious model. Even if parsimony-sensitive fit indices are used, a researcher might argue that additional parameters are strongly justified only if they substantially improve fit of the model to the data. Simple theories are often preferable unless a more complicated theory can substantially improve understanding of the phenomenon or can substantially broaden the types of phenomena understood under the same conceptual umbrella.

Model modification. In many cases, a researcher's model fits the data poorly. In such situations, the researcher is tempted to look for ways to improve the model. In fact, many statistical programs include indices (e.g., the modification index, Lagrange multiplier test, or Wald test) that provide information regarding improvement in fit if parameters in a model are freed. In practice, researchers frequently acknowledge that model parameters have been changed from a theoretical model based on these empirical indices of possible model modification. However, examinations of modification indices have revealed them to be highly problematic (e.g., MacCallum, Roznowski, & Necowitz, 1992). Such indices have been found to produce poor cross-sample replication, and tests in simulated data have revealed that they frequently fail to identify model misspecifications correctly. Therefore, empirical indices of model modification should be used with great caution. Changes made on the basis of such indices should always be theoretically plausible and whenever possible should be validated in a second sample. Ultimately, a researcher may be better

served modifying a model solely on theoretical grounds.

Design Issues in CFA.

As in EFA, design issues can have a substantial impact on the value of the results obtained from CFA. The primary design issues in CFA are similar to those of EFA. That is, CFA requires researchers to attend carefully to the selection of measured variables and research participants. However, the shift from an exploratory to a confirmatory approach requires some change in emphasis.

Selection of measured variables. Because CFA is employed when a clear basis exists for postulating the number and nature of the common factors, a researcher should select the best available measured variables to represent each common factor, rather than being concerned about comprehensiveness of the items (as in EFA). Therefore, prior data and theory should guide selection of optimal indicators of each common factor. As in EFA, CFA is most effective when each factor is overdetermined (i.e., at least three to five measured variables substantially load on the factor). Selecting measured variables with good psychometric properties is even more important in CFA. As in EFA, CFA is likely to perform better if the measures have high reliability. However, it is particularly important in CFA that each measured variable be influenced substantially by the common factor(s) postulated to underlie that measure and that it not be substantially affected by common factors other than those specified to influence the measure.

Selection of research participants. As in EFA, a researcher should carefully consider sampling and the number of participants to be included. Issues of selective sampling in CFA are similar to those in EFA, and the quality of measured variables and number of measures for each factor are relevant when estimating needed sample size. When the measures have little unique variance and common factors are overdetermined, relatively accurate parameter estimates can be obtained with modest sample sizes.

However, because CFA usually involves more formal assessment of model fit and specific hypothesis tests, issues of statistical power are more relevant than in EFA. There are a variety of ways of conceptualizing statistical power in CFA and SEM more generally, depending on the sort of hypothesis to be tested. For example, Saris and Satorra (1993) have focused on power analysis using the likelihood ratio test to compare a model with an alternative assumed to be true in the population. In contrast, MacCallum, Browne, and Sugawara (1996) proposed

a power analysis approach based on the RMSEA measure of model fit. In this approach, one specifies a hypothesis regarding RMSEA, an assumed value of RMSEA in the population, and a desired level of power (e.g., $RMSEA \leq 0.05$, assumed population RMSEA of 0.08, power of .80). A researcher can then compute the sample size necessary to obtain the desired level of power given the assumed population value of RMSEA. The MacCallum and colleagues' approach does not necessitate specification of an alternative model assumed to be true, and can be extended to other hypotheses regarding RMSEA and to fit indices other than RMSEA (e.g., MacCallum & Hong, 1997). Power analysis has also been developed for testing various hypotheses comparing nested models (MacCallum, Browne, & Cai, 2006).

Multidimensional Scaling

Multidimensional scaling (MDS) refers to a class of geometric models used to understand the underlying structure of (dis)similarities among objects (or concepts). MDS is sometimes used when developing a typology for a particular domain of objects. For instance, Rusbult and Zembrodt (1983) used MDS to examine constructiveness-destructiveness and activity-passivity as two dimensions underlying a typology of reactions to relationship dissatisfaction. MDS is also used to identify underlying dimensions people use to differentiate objects. For example, Feldman (1995) used MDS to address the dimensions (arousal and valence) that underlie perceptions of various emotions. As with EFA, MDS is often used in an exploratory fashion when there is no strong basis for predicting the number and nature of the dimensions. Additionally, dimensions can be used as dependent measures, influenced by some independent variable. For instance, Halberstadt and Niedenthal (1997) examined the extent to which people in various moods used a valence (emotion) dimension in perceiving similarity between faces (see also DeSteno & Salovey, 1997).

MDS starts with a proximity matrix among objects. In most applications, these proximities are similarity ratings, but other forms of data such as correlation coefficients can also be used (Borg & Groenen, 2005; Kruskal & Wish, 1978). Next, a model is selected and fit to the proximity matrix to derive coordinates that specify the location of each object in multidimensional space. This fitting process involves finding sets of values that minimize the discrepancy (as defined by some mathematical function) between the distance reconstructed from the MDS model and the observed matrix of proximities. The resulting pictorial representation of the objects is then examined to interpret the dimensions and/or

to develop typologies of a category of entities (e.g., objects or situations).

Statistical Issues in Conducting MDS.

When using MDS, a researcher must (a) choose which type of model to use, (b) determine the appropriate dimensionality in which to represent the objects, and (c) evaluate the adequacy of the resulting representation and interpret its substantive implications.

Choice of MDS model. One distinguishing characteristic of MDS models is the manner in which distance in multidimensional space is defined (i.e., the “metric”; MacCallum, 1988). The most common metric is Euclidean, and the most prevalent model is the unweighted Euclidean distance model (classical MDS). This model locates objects in multidimensional space using coordinates equal in number to the number of dimensions in the space, and the distances among objects are defined as simple distances (which can be measured with a ruler). Although this metric is the most intuitive way to define distances, some methodologists have proposed that other metrics might be more appropriate for certain types of data. Probably the best known non-Euclidean metrics are the city-block metric and the dominance metric. All three are special cases of a family known as Minkowski-p metrics, in which different members of the family are defined by different power functions. Several empirical and interpretational criteria have been proposed for choosing among metrics (Borg & Groenen, 2005; MacCallum, 1988). It seems preferable to use the common and intuitively appealing Euclidean metric unless there is a compelling conceptual and/or empirical basis that it is inappropriate. Unfortunately, definitive studies of the robustness of Euclidean models to violations of assumptions regarding the metric of the data have yet to be conducted (Borg & Groenen, 2005).

A second distinguishing characteristic among MDS models is the assumption regarding levels of measurement. Metric MDS models assume interval-level proximities whereas nonmetric MDS models assume only ordinal proximities. This distinction concerns the proximity measures themselves and has nothing to do with properties of the solutions. In almost all cases, both types produce coordinates with at least interval-level properties. In practice, metric and nonmetric MDS usually produce similar solutions (Schiffman, Reynolds, & Young; 1981; Borg & Groenen, 2005). Nonetheless, the choice is sometimes consequential. Metric MDS offers greater resistance to local minima and degenerate solutions (two problems discussed later; Borg & Groenen, 2005; Kruskal & Wish, 1978). Nonmetric MDS has the advantage of less stringent

assumptions regarding measurement properties of data. Also, determining the appropriate dimensionality can be somewhat easier in nonmetric MDS (Kruskal & Wish, 1978).

A third distinguishing feature of MDS models is the “mode” of data. Mode refers to whether the model is appropriate for data from a single source (i.e., matrix) or multiple sources. One-mode MDS is used when a single matrix of proximities (e.g., a matrix of proximities from a single individual or a matrix of proximities aggregated across individuals) is analyzed. In two-mode MDS, matrices could be two or more sets of aggregated proximities from participants in different experimental conditions or from the same participants at different points in time. Alternatively, matrices can be from individual participants. The term “two-mode” is used because the data are said to have two modes (i.e., objects and sources). The best known of the two-mode MDS models is probably the weighted Euclidean MDS model, also known as individual differences scaling (INDSCAL). Weighted Euclidean MDS differs from one-mode MDS in several ways. One-mode MDS produces a single stimulus space that represents the location of objects. Weighted MDS produces a similar “group stimulus space,” but this group space is not necessarily appropriate for any individual matrix (e.g., because individuals likely use each dimension to differing degrees). A set of weights is computed and used to adjust the group stimulus space (i.e., dimensions are stretched or shrunk) to produce a “personal” stimulus space. This feature is particularly useful when hypotheses involve differences among groups from different experimental conditions or personality types. For example, DeSteno and Salovey (1997) investigated the extent to which the dimensions underlying self-concept were influenced by temporarily induced mood states (see also Halberstadt & Niedenthal, 1997).

Unfortunately, researchers sometimes inappropriately compare groups on INDSCAL weights for a single dimension. Because INDSCAL normalizes data from each subject separately, the data are treated as “conditional” (i.e., the observations on one individual are not comparable to observations on another individual; Takane, Young, & De Leeuw, 1977). As demonstrated by MacCallum (1977), direct comparison of subject weights from INDSCAL is inappropriate. Such problems can be corrected if INDSCAL weights are compared using weight ratios (i.e., ratios of use of one dimension compared to use of another dimension; MacCallum, 1977; Schiffman et al., 1981).

A second important difference between one-mode and weighted MDS regards orientation of the dimensions. In one-mode MDS, dimensions can be rotated

without affecting the ability of the representation to account for proximities. Thus, a researcher must determine the most useful orientation. In weighted MDS, there will be one orientation that best accounts for proximities. Although a unique solution does not ensure that this orientation will have conceptual meaning, empirically this is usually the case (Wish & Carroll, 1974).

Determining dimensionality. After selecting a model, a researcher must determine the appropriate dimensionality with which to represent the objects of interest. This task is similar to selecting the appropriate number of factors in EFA. Like EFA, dimensionality in MDS is as much a theoretical question as a statistical question (Kruskal & Wish, 1978). Statistical procedures exist to aid in this decision, but the issue of interpretability and theoretical plausibility must always be taken into account (MacCallum, 1988).

One approach is to conduct a series of MDS analyses beginning with a one-dimensional solution and adding a dimension with each analysis until a solution includes three more dimensions than is expected to be needed (Harshman, 1984). A plot of the fit of each solution (as assessed by one of the fit indices discussed later) is then constructed to determine when adding dimensions no longer produces a substantial improvement in fit. This approach is analogous to use of goodness of fit indices in determining the number of factors in EFA (see earlier discussion). As in EFA, identifying breaks in improvement of fit can be rather subjective.

Another approach is to compare the plot of goodness of fit for a series of analyses with plots of fit that would be expected to occur for data with varying numbers of underlying dimensions (Kruskal & Wish, 1978). Methodologists have compiled numerical values of goodness of fit that would be obtained if the true underlying dimensionality were one-dimensional, two-dimensional, and so forth. These plots can be examined to see which plot most closely matches the plot generated from the actual data. Presumably, the underlying dimensionality is the same as the dimensionality of the best-matching expected plot. Unfortunately, plots of expected fit have not been compiled for many types of models and indices of fit. Also, these plots reflect dimensions of relatively equal importance (Kruskal & Wish, 1978), which may not be the case with real data.

A third basis for selecting dimensionality is the stability of solutions at different dimensionalities (Kruskal & Wish, 1978; MacCallum, 1988). Stability is assessed by examining the solution for randomly split halves of the sample or by assessing changes in the solution when an object is deleted from the analysis. Thus, a researcher can examine the point at which introducing additional

dimensions reduces the stability of the solution. Only a solution in which all dimensions are stable should be used as a basis for interpretation.

Evaluating and interpreting MDS solutions. As alluded to earlier in our discussion of determining dimensionality in MDS, various fit indices have been developed to assess MDS models (for a review, see Borg & Groenen, 2005). Most fit indices are of a form known as *stress*. Stress is a numerical value reflecting how poorly the model fits the data. Thus, smaller values indicate better fit. Two of the best-known indices of stress are Stress 1 and S-stress. Although stress is the most common method of expressing model fit in MDS, nonstress indices are also sometimes used (e.g., RSQ in weighted Euclidean MDS; Schiffman et al., 1981). If an MDS analysis indicates poor fit, the results should be interpreted with caution. Guidelines have been proposed for some indices (e.g., Kruskal, 1964; Takane et al., 1977), but fit indices can be affected by a number of factors (Borg & Groenen, 2005).

When evaluating solutions, researchers must also be attentive to the possibility of local minima and degenerate solutions. A local minimum occurs when a MDS procedure fails to converge on a best-fitting solution. This problem can be detected by examining a plot of stress values at differing dimensionalities. If increasing dimensionality is found to produce a larger stress value, this suggests the existence of a local minimum (Kruskal & Wish, 1978). One could also conduct several analyses for the same dimensionality and data using different initial start values for the parameter estimates (Borg & Groenen, 2005; Harshman, 1984). If different start values produce the same stress value and parameter estimates, a local minimum is unlikely. A degenerate solution refers to a situation in which the index of fit can be made arbitrarily small regardless of the relation between the proximity data and the inter-object distances in the MDS solution (Borg & Groenen, 2005). Degenerate solutions can generally be avoided by placing a sufficient number of constraints on the model (Borg & Groenen, 2005).

Excellent discussions of substantive interpretation of MDS solutions are available (Kruskal & Wish, 1978; MacCallum, 1988; Schiffman et al., 1981). One subjective technique is to examine the objects located at the extreme ends of each dimension. The dimension is interpreted by identifying characteristics common to all objects at the same end of the dimension. A more formal approach uses regression analysis. A researcher first identifies characteristics of the objects that might constitute underlying dimensions, and collects ratings of each object on these characteristics along with the similarity judgments among

objects. Dimension coordinate values for each object are used as predictors in a regression, and ratings of the characteristics for each object are used as DVs. Dimensions found to be particularly strong predictors of a characteristic are implied to reflect that characteristic (e.g., see Rusbult, Onizuka, & Lipkus, 1993 for an illustration of this approach in the context of romantic relationships).

Design Issues in MDS.

A number of design features should be taken into account when collecting MDS data. Perhaps the most important issue is the selection and content of objects. For example, the stress of a model can be misleading when the number of objects relative to dimensions is low. Therefore, it is important to consider how many dimensions are expected to emerge and to ensure that a large number of objects relative to dimensions is included (i.e., at least 4:1; Kruskal & Wish, 1978). Also, in any MDS analysis, the dimensions that emerge depend on the objects that are included in the analysis. If a domain is inadequately sampled, important dimensions could fail to emerge. It is also important not to include irrelevant objects because results can be distorted by large differences between relevant and irrelevant objects.

Another design issue is the manner in which proximity data are collected. When direct similarity judgments are collected, a variety of possible judgment procedures can be used (Borg & Groenen, 2005; Schiffman et al., 1981). The researcher must also consider controlling for order effects in judgments and whether to include ancillary measures of objects to be used in regression analyses to aid in interpretation of dimensions. In some cases, the number of objects might be so great that it is not feasible to collect similarity ratings for all possible pairs. Procedures have been developed for collecting and analyzing data for incomplete proximity matrices (Borg & Groenen, 2005; MacCallum, 1988; Schiffman et al., 1981). Sometimes direct similarity judgments might not be available and thus data must be converted into a form of proximity data (see Borg & Groenen, 2005).

A final issue is the selection of research participants. Samples should, when possible, represent the population of interest. It is also useful to measure individual differences on constructs related to the manner in which participants judge the objects of interest. For instance, Carroll and Wish (1974) found that individuals' characteristics such as age, gender, political ideology, and religion influenced use of dimensions in similarity ratings of different types of interpersonal relationships. Similarly, one could measure individual differences

such as affect intensity and investigate whether such differences relate to dimension use in affect-relevant judgment (e.g., see Feldman, 1995; Halberstadt & Niedenthal, 1997).

Summary and Comparison of Noncausal Methods

We have described three procedures for addressing noncausal hypotheses regarding the number and nature of dimensions underlying a set of data: EFA, CFA, and MDS. Although these procedures are often treated as distinct approaches, it should of course be recognized that there are a number of underlying mathematical similarities. We have already noted the many similarities between EFA and CFA. However, there are also many underlying mathematical relationships between MDS and EFA (for a detailed discussion of this issue, see MacCallum, 1974). That being said, the methods are to some degree distinguished by the types of data they typically utilize and the types of questions they are commonly used to address. Whereas EFA and CFA investigate relations among measures of a given object, MDS is more commonly used to address the dimensionality of a set of objects. Some research questions lend themselves to ratings of a given object, whereas others lend themselves to arrangement/judgment of multiple objects (especially on similarity). Some form of factor analysis is typically well suited to the former type of question, whereas MDS is well suited to the latter type of question. It is also important to note, however, that there might often be opportunities for addressing the same conceptual question from either approach (and there is nothing to say that one could not “factor analyze objects” by rating many objects on a single dimension or use MDS to spatially represent a set of traits relevant to a single object, although these are rarely done).

When a research question seems equally amenable to either strategy, we believe that there are reasons to favor factor-analytic methods. One reason is that the common factor model explicitly represents measurement error within the model. Thus, the existence of measurement error can to some degree be assessed and accounted for in the solution. MDS models do not incorporate measurement error, so it is difficult to determine the severity or impact of such error in a particular data set. A second advantage of factor analysis regards rotation. Although the orientation of dimensions in weighted Euclidian MDS is uniquely determined, this is not true in most other forms of MDS, so that rotation becomes an issue. In factor analysis, criteria and specific procedures for both orthogonal and oblique rotation have been well developed and are available in

software. This is less true for MDS. Another major advantage of factor analysis is the fact that ML estimation has been extensively studied and is available in virtually all major statistical programs. This allows for the computation of a wide range of indices of goodness of fit and provides the ability to compute confidence intervals and significance tests for parameter estimates. Although ML procedures have been developed for MDS (e.g., Borg & Groenen, 2005), they have not been as extensively studied, nor are they as widely available. Finally, confirmatory approaches to factor analysis have been extensively developed and computer software for implementing CFA is widely distributed. In contrast, although there is some work on confirmatory MDS (e.g., Borg & Groenen, 2005), this literature is much less developed and software is not as widely available.

There are certainly some situations in which MDS is desirable. These would include research questions that explicitly deal with dimensions along which people classify objects or with ways in which certain individuals or groups use particular dimensions to classify or perceive objects. Also, MDS does not require explicit specification of the dimensions along which objects vary, unlike EFA and CFA. However, when none of these research goals or concerns is central, EFA or CFA may be preferable.

EFA, CFA, and MDS are certainly not the only procedures that address noncausal hypotheses. We focused on these procedures because they have been extensively used within social-personality psychology. However, other techniques addressing similar types of questions may also potentially be of interest to social-personality psychologists. Of these perhaps the best known are cluster analysis (Everitt, Landau, Leese, & Stahl, 2011), correspondence analysis (Greenacre, 2007), and item response theory (IRT; Embretson & Reise, 2000; see also Widaman & Grimm, Chapter 20 in this volume).

Analyses Involving Causal Hypotheses

In many research settings, the key questions of interest go beyond determining the number and nature of constructs underlying a set(s) of measured variables or objects. In many settings, a researcher might wish to examine potential causal relations among the constructs. We begin our discussion with an overview of the major types of causal hypotheses. Then we discuss the conditions necessary for establishing causal relations and comment on study design features and statistical procedures that assist in establishing these conditions. Finally, we

review statistical procedures used to test different types of causal hypotheses. Our review focuses on procedures that have traditionally been associated with the analysis of nonexperimental data. However, these procedures can fruitfully test causal hypotheses with either experimental or nonexperimental data (see also Judd et al., Chapter 25 in this volume; Smith, Chapter 3 in this volume).

Types of Causal Hypotheses

The simplest form of causal hypothesis is that a given IV (measured or manipulated) is hypothesized to have a direct causal influence on a DV. For example, a researcher might hypothesize that increases in the value of an IV directly lead to increases in the value of a DV. We refer to such hypotheses as *direct causal* hypotheses. Obviously, researchers might often be interested in testing more than one direct causal hypothesis. One might hypothesize that multiple IVs directly influence a DV, that one IV directly influences multiple DVs, or that multiple IVs directly influence multiple DVs. In some cases a researcher might further expand on these direct causal hypotheses by postulating a “relative strength of direct causal relations” hypothesis. Such a hypothesis might involve postulating that the effect of IV1 is greater than that of IV2 on a particular DV, or that the effect of an IV is greater on DV1 than it is on DV2.

In other cases, a researcher might hypothesize a more complex set of causal relations among one or more IVs and one or more DVs (see also Judd et al., Chapter 25 in this volume). One such type of hypothesis is a mediational hypothesis (Baron & Kenny, 1986). Mediation refers to a case in which a researcher postulates that an IV exerts a causal influence on a DV at least in part indirectly, via its influence on another (mediating) variable. That is, a researcher postulates that an IV has a direct causal influence on a mediator variable, which in turn has a direct causal influence on the DV. Mediational hypotheses are often complex. For example, a researcher might postulate that more than one variable mediates the impact of the IV on the DV. Alternatively, the researcher might be interested in multiple IVs or multiple DVs. Finally, a researcher might posit multiple steps in a mediational chain.

A third type of causal hypothesis, moderation (Baron & Kenny, 1986), refers to a case in which the strength and/or valence of a relation between an IV and DV is thought to be regulated by a second IV (the moderator). For example, a researcher might hypothesize that the first IV will have a substantial influence on the DV at one level of the second IV but that this influence will become weaker at another level of the second IV. Alternatively, the researcher might predict that

the first IV will have a positive influence on the DV at one level of the second IV but that this influence will be negative at another level of the second IV.⁷ Such moderator hypotheses are typically examined using interaction tests among IVs. Complex moderator hypotheses are also often examined. For instance, a researcher might hypothesize that a moderator relation between two IVs is moderated by a third IV (i.e., a three-way interaction).

Although more rarely discussed, many theories also involve “hybrid” hypotheses that combine different types of causal hypotheses (see also Judd et al., Chapter 25 in this volume). For instance, a researcher might hypothesize that the relative impact of an IV on two different DVs varies across levels of some variable (i.e., moderated relative impact), or that a relative difference in the impact of two IVs on a DV varies across levels of another independent variable.

In some cases, a researcher might postulate a mediational relation that varies across levels of some variable; that is, one variable moderates the mediational relations among a set of variables – *moderated mediation*. This might involve different variables mediating the relation between an IV and a DV at different levels of the moderator. For example, the Elaboration Likelihood Model involves moderated mediation (Petty, Wegener, Fabrigar, Priester, & Cacioppo, 1993), with assessments of the central merits of attitude objects mediating the effects of persuasive appeals under conditions of high elaboration, but with simplified processes (e.g., heuristics or conditioning) mediating the effects of the same appeals under conditions of low elaboration. Alternatively, it might be hypothesized that an IV has a direct impact on a DV at one level of the moderator, but this relation is mediated at another level of the moderator.

Another form of hybrid hypothesis occurs when an IV moderates the impact of a second IV via its influence on some mediator variable – *mediated moderation* (Muller, Judd, & Yzerbyt, 2005). For example, a moderator might causally influence a mediator, which in turn moderates the impact of another IV on the DV. Such a hypothesis is implied by most moderation effects in social psychology. That is, a manipulation's effect on a psychological variable (often measured as a manipulation check) is hypothesized as responsible for the moderational impact of the manipulation. Of course, other examples exist as well. For instance, DeSteno, Petty, Wegener, and Rucker (2000) found that different negative emotions (i.e., anger vs. sadness) produced opposite patterns of likelihood judgments for angering versus saddening events (a moderation effect). DeSteno and colleagues went on to test the possibility that the moderating effect of emotions was mediated by global beliefs about whether the

world was a maddening versus saddening place.

There is also another type of mediated moderation in which there is a moderational effect of two IVs on the mediating variable and a direct effect of that mediator on the DV (Baron & Kenny, 1986). This type of mediated moderation might be common in studies where the DV is some type of behavior. For example, one might find that choices of products are influenced by the content of product advertisements when the product appears to be personally relevant to consumers, but are not when the product appears low in personal relevance to consumers (e.g., Petty, Cacioppo, & Schumann, 1983). This moderational (content X relevance) effect on behavior, however, might take place because the content X relevance moderation influenced attitudes (e.g., Petty, Cacioppo, & Goldman, 1981) and attitudes influenced behavior (either directly or through behavioral intentions; Fazio, 1990).

A final type of hybrid hypothesis occurs when noncausal relations among constructs or variables are moderated by some variable. For instance, a researcher might postulate that a moderator variable (causally) influences the magnitude of correlations between two common factors underlying a set of measured variables. Alternatively, a researcher might hypothesize that the number of common factors differs across levels of a moderator variable or that the magnitude of factor loadings differs across levels of a moderator variable.

Conditions for Inferring Causality

These various hypotheses all share the common feature of implying at least some form of causality among the constructs of interest, which can be represented mathematically. Some statistical procedures permit representation of virtually any of these types of hypotheses, whereas other procedures allow for representation of only some of the hypotheses. However, it is important to recognize that ability to use a statistical procedure to represent a hypothesized set of causal relations does not necessarily imply that the hypothesized causal relations are correct. For example, a researcher can conduct a regression analysis assessing the impact of several IVs on a DV. Although this analysis models a directional impact of the IVs on the DV, finding significant effects of the IVs certainly does not establish that the relations are causal.

A substantial literature addresses the conditions under which causality can be inferred (e.g., Judd & Kenny, 1981; see West, Cham, & Liu, Chapter 4 in this volume). Although such discussions vary, three commonly noted conditions are isolation, association, and direction (e.g., Bollen, 1989). A DV can only be said

to be caused by an IV when the impact of the IV on the DV has been isolated from all other influences (i.e., there are no other potential IVs that might account for the effects of this IV on the DV). It must also be established that changes in the IV are associated with changes in the DV. Finally, the direction of the association must be that changes in the IV lead to changes in the DV rather than the reverse. Establishing the existence of these conditions can be challenging, and there are a number of features of both study design and statistical analysis that affect this determination (see Brewer & Crano, Chapter 2 in this volume; Smith, Chapter 3 in this volume; West et al., Chapter 4 in this volume).

The conditions of isolation, association, and direction can be thought of as falling along a continuum. At one end of the continuum is simple measurement of the IV and DV, with no attempts to control for extraneous influences on the DV. Even if an association between the IV and DV is found, no claims about isolation or direction are possible. Statistical techniques that control for the impact of alternative IVs, for measurement error, or for both can move one higher on the continuum toward building a credible case for the causal effect of an IV on a DV. Statistical techniques alone, however, are less effective than are design features such as random assignment to conditions of an experimentally manipulated IV. Although causal inferences based on experimental manipulations pose additional issues, common design features of experiments can at least indicate that the manipulation itself causes a change in the DV.

In social psychology, experiments have become the standard approach for addressing causal hypotheses. However, as noted earlier, there are many reasons for conducting nonexperimental research. Given nonexperimental data, some statistical methods are more useful than others for establishing at least a modest level of confidence that the conditions for inferring causality have been satisfied. In the following sections, we review several statistical procedures that are most commonly associated with testing causal relations in nonexperimental settings.

Regression

Multiple linear regression has, over the past 20 years, become a common, “everyday” method of analysis for both experimental and nonexperimental research in psychology. Regression is a general and flexible technique, capable of addressing many types of research questions involving both continuous and categorical predictor variables. In general, linear regression consists of a set of predictor variables (IVs) hypothesized to influence a single DV. These IVs can be continuous or categorical measures of the constructs of interest and other

relevant variables; DVs are typically continuous measures. Regression IVs are related to the DV through a set of linear equations (Cohen, Cohen, West, & Aiken, 2003).

As presented in Figure 19.4, multiple regression models are saturated, with all IVs correlated with one another and all IVs influencing the DV. In addition, the IVs and DV are treated as perfect measures of the constructs of interest. That is, there is no representation of measurement error. It is therefore impossible to separate errors in equations (predicting the DV from the IVs) from errors of measurement (see later discussion of SEM). The saturated nature of the model, with measures treated as perfect indicators, render the model nonfalsifiable (i.e., the model exhausts all degrees of freedom, such that model fit per se cannot be assessed). Thus, the focus is generally on parameter estimation to determine the extent to which the various IVs uniquely influence the DV (controlling for the impact of other predictor variables).⁸

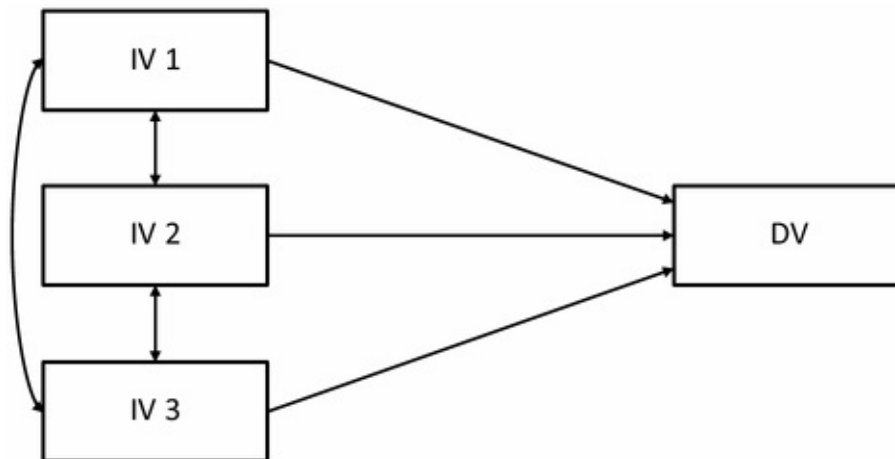


Figure 19.4. Path diagram of the multiple regression model for direct causal hypotheses.

Types of Hypotheses.

Direct cause. The model represented in Figure 19.4 tests direct causal hypotheses. That is, the IV(s) of interest directly influence the DV, controlling for other potential influences. If a researcher wishes to distinguish the IV(s) of interest from related, but conceptually distinct, influences on the DV, these alternative predictors can be measured and included in the model. Of course, as was noted earlier, the fact that the model assumes such casual relations between IVs and the DV does mean that these causal assumptions are correct. However, researchers should be aware that such assumptions are implicit anytime a

regression analysis is conducted.

Relative strength of direct causes. Researchers using regression often reach conclusions regarding the relative impact of different IVs on a given DV or a particular IV across two or more DVs. However, actual statistical tests in support of such conclusions are much rarer than one might expect. For example, researchers sometimes examine regression models with multiple predictors and, based on the resulting regression coefficients, conclude that some IVs have stronger effects on the DV than do others, because some coefficients are statistically significant whereas others are not, or because some of the coefficients appear to be bigger than others. Strictly speaking, concluding that two coefficients are different from another within a model (in the case of comparing two IVs) or across models (in the case of comparing an IV's effects across two DVs) cannot be done without formally testing the difference between these coefficients.

It is possible to conduct such tests. For example, if two IVs are scaled to a common metric (i.e., the two IVs are on the same scale of measurement), comparison of their effects on a given DV is straightforward from a computational standpoint (Judd & McClelland, 1998). The two IVs being compared are summed and a second regression analysis is run that is identical to the original analysis, except that the two IVs are replaced by the single summed variable. This model is mathematically equivalent to a model constraining the coefficient for the two IVs to be equivalent. The R^2 for the new model is then compared to the R^2 of the original model and an increment in R^2 test is conducted (Cohen et al., 2003; Judd, McClelland, & Ryan, 2008). If the new model accounts for significantly less variance than does the original model, this indicates that the constraint was not realistic and thus the two coefficients are significantly different from one another. Alternatively, a researcher might hypothesize that an IV has a stronger influence on one DV than on another DV. This could be tested by creating a difference score between the two DVs (assuming the two variables share the same scale of measurement) and using the IV to predict the difference score. If the IV significantly predicts the difference score, this is equivalent to testing the difference in regression coefficients relating the IV to each individual DV.

Although it is easy to compute test statistics for comparisons of relative strength within regression, interpreting differences can be much more challenging. For example, there are several reasons that one IV might have a greater impact than another. Such a difference might emerge because the

underlying construct for one IV is more consequential than the construct underlying the second IV. However, within regression, differences might also emerge because there is simply more variance in a given sample for one IV than the other or because the measure of one IV is more reliable or valid than that of the other. Likewise, in the context of experimental manipulations, the manipulation of one construct might simply be more successful than the manipulation of the other. In comparisons across DVs, once again differences could emerge because the two constructs represented by the different DVs really are differentially affected by the construct reflected by the IV. However, differential variance in the DVs (which could occur artificially if the DVs are on different scales) and differences in the reliability or validity of the measures are possible alternative explanations. Moreover, when IVs and/or DVs are originally in different scales of measurement, comparisons can be even more challenging. Standardization is potentially problematic because standardized coefficients are affected by the variability of the IVs and/or DVs. Some sort of other linear transformation that does not involve standardization can of course be used (e.g., scaling variables in question to a 0 to 1 metric), but this does not address the issue that the original scale format may have had an impact on the performance of the variable in the analysis rather than the construct it is intended to reflect.

Mediation. Regression can also address mediational relations among variables. For example, [Figure 19.5](#) presents a simple mediation, with the effect of the IV on the DV mediated by the mediator variable. Often such models also allow for direct influence of the IV on the DV. Use of regression to test such mediational hypotheses has become commonplace in social-personality psychology. Thus, we will not provide a detailed discussion of the mechanics of such analyses (see Judd et al., Chapter 25 in this volume; MacKinnon, [2008](#)). Briefly stated, the parameters of this model are estimated from two regression models: a model in which the IV influences the mediator and a model in which both the mediator and the IV influence the DV. If these models show significant relations between the IV and mediator, and between mediator and DV, this is consistent with a hypothesis of mediation. The most common statistical test of the indirect (mediational) effect is the Sobel test (Baron & Kenny, [1986](#)). However, this test lacks power, and the distribution of indirect effects (represented by the multiplication of the IV to mediator path and the mediator to DV path) routinely departs from normality. One way to improve power and address these distributional difficulties is to test the indirect effects using bootstrapping (Shrout & Bolger, [2002](#); for SAS and SPSS syntax, see Preacher & Hayes, [2004](#)).

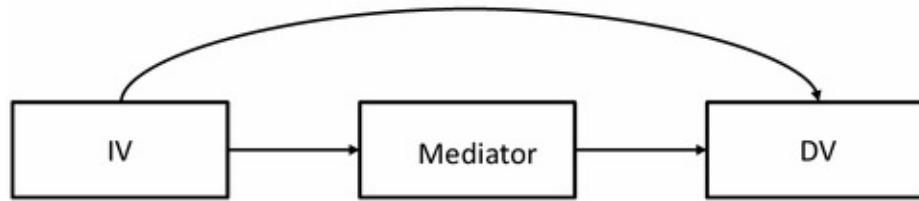


Figure 19.5. Path diagram of a simple mediational relation using measured variables.

Although regression remains the most commonly used method of assessing mediation, this approach has drawbacks. First, as the number of mediators increases, regression-based tests quickly become unwieldy (although bootstrapping macros have been developed for both multiple possible mediators operating in parallel; Preacher & Hayes, 2008; and for multiple sequential mediators; Hayes, Preacher, & Myers, 2011). Second, the approach does not take into account measurement error, which can result in overestimation of direct effects of the IV on the DV. Thus, when feasible, the regression technique for testing such hypotheses is best abandoned in favor of analyses of covariance structures (see later discussion).

Moderation. One strength of the regression approach is the ease with which it deals with moderational relations. As depicted in Panel A of Figure 19.6, moderators are often tested by including a multiplicative term in the model (i.e., a product of the two interacting IVs). The interaction is tested as the unique influence of the multiplicative term when the lower-order (main effect) IVs are included in the model. Many excellent discussions of moderated regression are available, (i.e., Cohen et al., 2003; Judd et al., Chapter 25 in this volume).

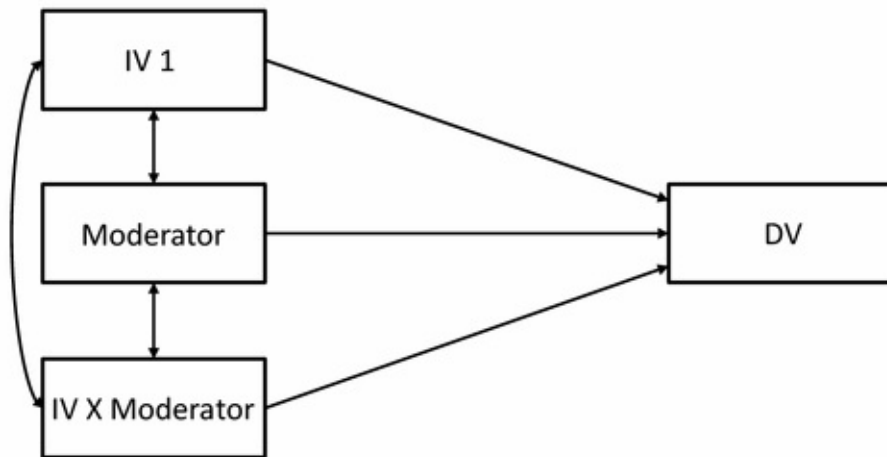
PANEL A**PANEL B****Moderator Group 1****Moderator Group 2**

Figure 19.6. Path diagram of tests of simple moderational relations in multiple regression.

When a moderator variable is dichotomous or categorical, one can address moderational hypotheses by running separate regressions at each level of the moderator variable. As shown in Panel B of Figure 19.6, one might conduct separate analyses investigating the impact of an IV on the DV at each level of the moderator variable, then testing the difference in regression coefficients across pairs of groups (Cohen & Cohen, 1983). Although such a test is not identical to the interaction test, the test of the difference between these coefficients across the groups does address the conceptual question of difference in impact of an IV across levels of the moderator. Such an approach is rare for

straight moderation tests (and would often be less powerful than the traditional test), but the split analysis becomes more useful when attempting to address some of the hybrid hypotheses using regression.

Moderated mediation. Moderated mediation, especially when the moderator is categorical, has most commonly been examined by separate mediational analyses at each level of the moderator. Then one can conduct comparisons across levels of the moderator between the regression coefficients for each of the three critical paths (i.e., IV to mediator, mediator to DV partialing the IV, and IV to DV partialing the mediator). In some cases, moderated mediation might be caused primarily by changes in impact of the IV on the mediator, in others by changes in the impact of the mediator on the DV, and in yet others by both. Moderated mediation could occur when a moderator \times IV interaction is observed on the DV (because of differences in IV to mediator and/or mediator to DV paths) or when no moderator \times IV interaction is observed on the DV (because different mediators create the same magnitude of effect or a mediator operates at some levels of the moderator but direct effects occur at other levels; e.g., Wegener, Clark, & Petty, 2006). Recently, detailed discussions have appeared of approaches using regression models with interaction terms to identify moderated mediation effects (Judd et al., Chapter 25 in this volume; Muller et al., 2005; Wegener & Fabrigar, 2000; see Preacher, Rucker, & Hayes, 2007 for bootstrapping methods to test moderated mediation and mediated moderation).

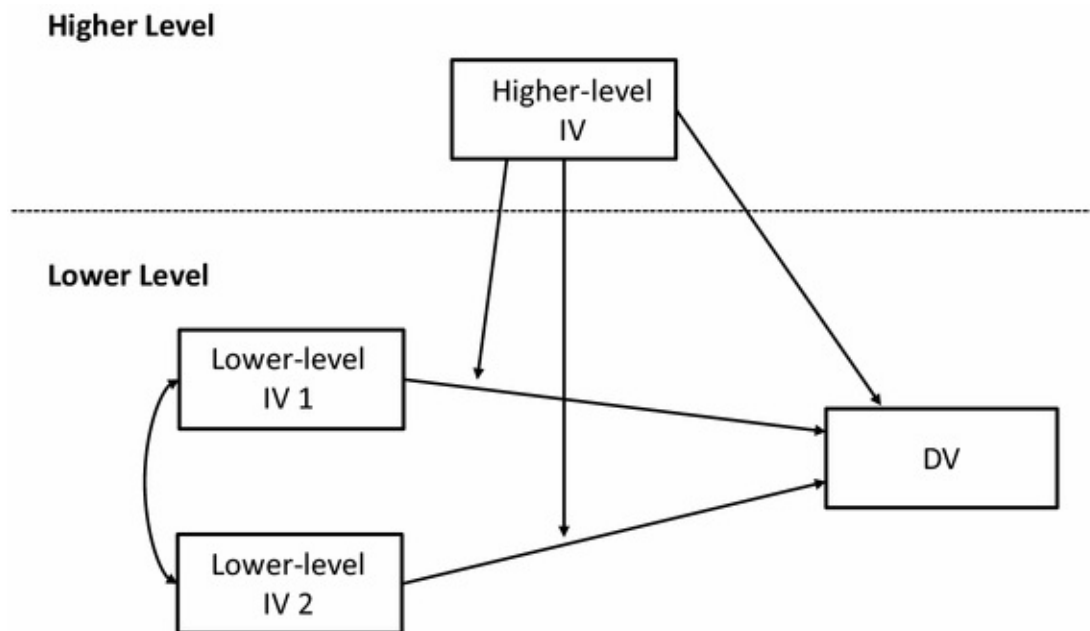


Figure 19.7. Simplified representation of two-level HLM. *Note:* Lower-level (level 1) variables are presumed to be nested within higher-level (level 2) variables

Mediated moderation. In other contexts, researchers might be interested in mediated moderation – for example, when a moderator interacts with an IV to affect a DV, but the moderator has its effect via some mediating variable (which is the conceptual variable that actually interacts with the IV). As noted earlier, this logic is at least implicit when researchers collect manipulation-check data. The manipulation is assumed to affect a conceptual variable (measured by the manipulation check), which is assumed to interact with another IV. Interestingly, although use of manipulation checks has been common in social psychological research, tests of mediated moderation using such measures have been comparatively rare. However, such questions can be explored using regression-based analyses. When terms involving the manipulation check reduce the significance of parallel terms involving the manipulation (while retaining significant effects of the manipulation-check terms), the pattern of results is consistent with mediated moderation (Muller et al., 2005; e.g., Wegener, Petty, & Smith, 1995). Another form of mediated moderation occurs when an IV and moderator interact to affect a mediator, and that mediator then directly influences a DV. In such analyses, the mediator influences the DV above and beyond distal effects of the IVs, and including the mediator in the model reduces significance of the distal IV \times moderator interaction (Judd et al., Chapter 25 in this volume; Muller et al., 2005; e.g., see Clark, Wegener, & Fabrigar, 2008).

Multilevel Models

Data sometimes involve more than one level of analysis, where each lower level is nested within a higher level (Schoemann, Rhemtulla, & Little, Chapter 21 in this volume). For example, imagine that a researcher is interested in the determinants of participation in group discussions. One might obtain lower-level measures of characteristics of individual group members (e.g., knowledge of the discussion topic). These individuals are nested within groups (i.e., there is person-based variance within a given group), and the researcher might also obtain higher-level measures of characteristics of the groups (e.g., group size). The researcher could then test if knowledge and group size influence participation in the discussion.

One increasingly popular method for examining nested data is multilevel modeling (MLM; also called hierarchical linear modeling ; Raudenbush & Bryk, 2002; Schoemann et al., Chapter 21 in this volume). MLM can be thought of as a multilevel form of multiple regression, in which ML estimation is typically used (see earlier discussion of ML estimation in CFA). At the lower level of analysis,

the model is similar to traditional multiple regression, with a set of IVs predicting a DV (see [Figure 19.7](#) for a simplified representation). In our example, each member's topic knowledge would be an IV (along with any additional IVs) predicting that person's discussion participation. MLM models differ from traditional regression models in that a second level of analysis is also represented and in that observations at the lower level of analysis can be grouped together based on their membership in some higher-order level of organization (lower-level variables are nested within higher-level variables). For instance, in our example, individuals can be grouped because they belonged to the same discussion group. Furthermore, each set of lower-level observations can be conceptualized as having its own parameter estimates. Thus, within each discussion group, it is possible to compute regression coefficients and an intercept for the lower-level IVs predicting discussion participation. Importantly, the model also allows specification of IVs measured at the second level of analysis that are presumed to account for variations across sets of observations (i.e., groups) in the value of the lower-level regression coefficients and intercepts. For instance, group size could be specified as a higher-level IV predicting variation across discussion groups in the regression coefficients and intercepts for the lower-level IVs influencing discussion participation. Although two-level models are the most common, it is possible to test models with more than two levels.

At the lower level of analysis, the same types of causal relations that can be examined in multiple regression can be explored in MLM. Higher-level IVs accounting for variations in the lower-level regression coefficients can be conceptualized as moderator relations. That is, the effect of a higher-level IV on lower-level regression coefficients indicates that the influence of the lower-level IV (e.g., topic knowledge) on the DV (e.g., discussion participation) is not the same across different levels of a higher-level IV (e.g., group size). The test of whether a higher-level IV accounts for variations in the intercept of the lower-level model can have somewhat different conceptual implications depending on the scaling of the lower-level IVs. If the lower-level IVs are expressed as deviations from the mean value for that higher-level unit, an effect of a higher-level IV on the lower-level intercept corresponds to a direct causal relation between the higher-level IV and the DV (e.g., group size influences amount of member participation). In [Figure 19.7](#) this effect is depicted using a direct arrow from the higher-level IV to the DV.

MLM can also be used for testing more complex relations. For instance, one might imagine using MLM to test moderated mediation (Bauer, Preacher, &

Gil, 2006). A higher-level variable could moderate a set of mediational lower-level relations. In fact, such moderated mediation questions fit naturally into the logic of the model. For example, individuals could be considered “grouped” by the fact that they all received the same manipulation in an experiment (i.e., experimental condition is the higher-level IV). Effects more complex than two-way interactions can also be addressed. One might also imagine testing whether the moderational impact of one higher-level IV on the lower-level regression coefficients or intercepts is mediated by another higher-level IV (i.e., mediated moderation).

MLM has several advantages over traditional regression models for analyzing multilevel data, including improved estimation of lower-level effects within higher-level units, the ability to test cross-level effects, and the ability to partition variances and covariances in lower-level measures into within-unit and between-unit higher-level components (Raudenbush & Bryk, 2002; Schoemann et al., Chapter 21 in this volume). However, MLM is not the only way to deal with multilevel data. Traditional multiple regression models can be adapted (Kenny et al, 1998), and certain CFA models are mathematically equivalent to some classes of MLM (e.g., MacCallum, Kim, Malarkey, & Kiecolt-Glaser, 1997). Moreover, multilevel Structural Equation Models have been developed to address mediation in multilevel data (e.g., Preacher, 2011; Preacher, Zhang, & Zyphur, 2011; Preacher, Zyphur, & Zhang, 2010). Additional work remains to be done, but MLM has proven to be a very useful approach to the analysis of multilevel data. For psychology, the MLM method has proven especially useful in longitudinal studies of change (Schoemann et al., Chapter 21 in this volume) and settings for which nonindependence of observations is an issue, such as dyadic and group interactions (Kenny & Kashy, Chapter 22 in this volume)

Structural Equation Modeling

Structural equation modeling, or SEM (also called covariance structure modeling), is a method of specifying and estimating relations between latent variables and measured variables and among latent variables (Bollen, 1989; Schumacker & Lomax, 2010). SEM is a general mathematical framework that includes CFA (as well as regression, path analysis, and ANOVA) as a special case. One can think of SEM as a CFA model in which causal relations are permitted among the common factors. For example, the SEM depicted in Figure 19.8 includes three common factors in which IV1 and IV2 are correlated and each exerts a causal influence on the DV.

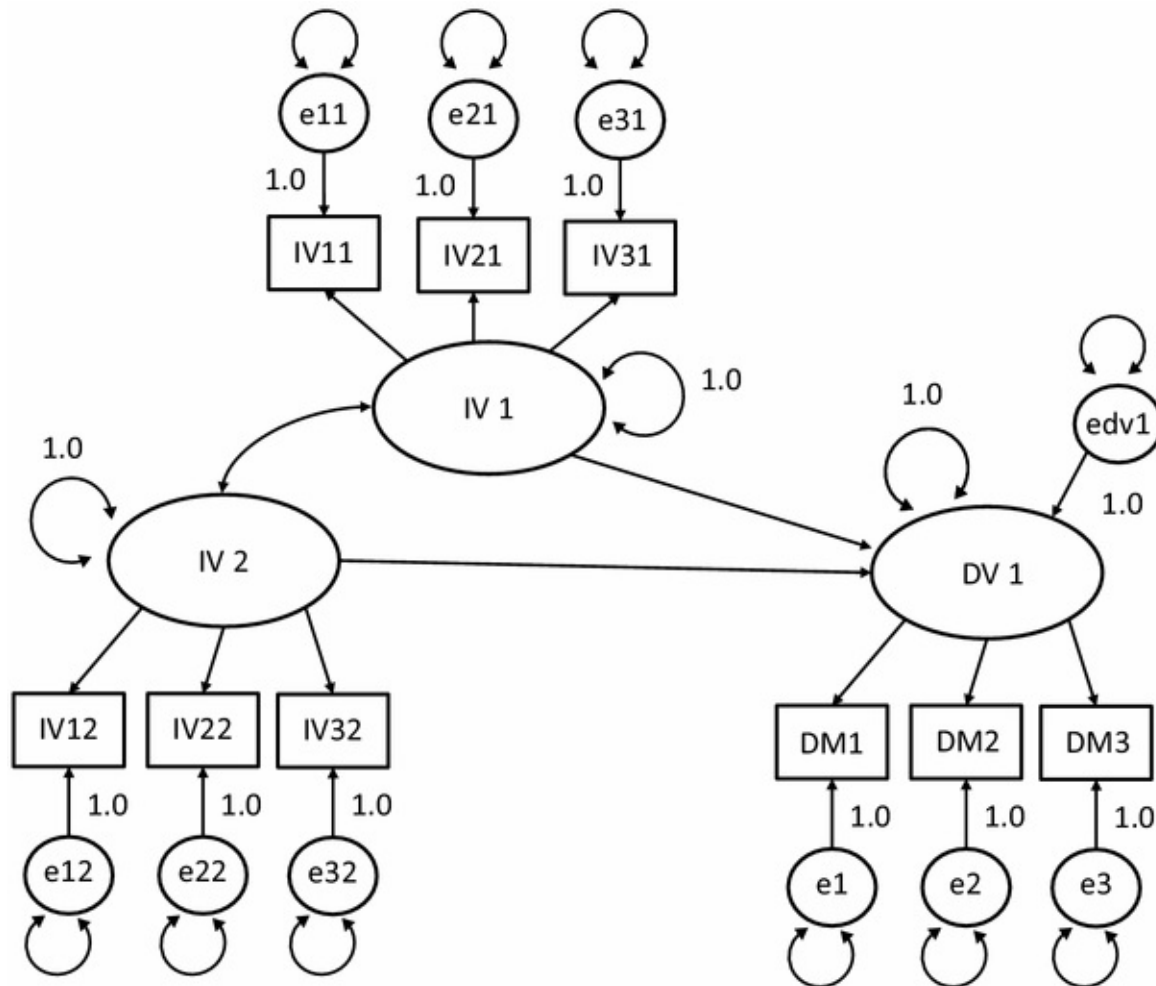


Figure 19.8. Path diagram of a SEM specifying two correlated latent variables that causally influence a third latent variable. *Note:* Latent variables are represented by ellipses and measures are represented as rectangles. Unique variances are displayed as latent variables using circles or ellipses. Directional relations between variables are represented as single-headed arrows and nondirectional relations as doubleheaded arrows. Variances of latent variables are represented by doubleheaded arrows from the variable to itself. Arrows without numerical values are free parameters, and numerical values of the arrows, are values of fixed parameters specified by the researcher.

Statistical Issues in SEM

In most respects, conducting an SEM analysis is very similar to conducting a CFA. Thus, we will only briefly touch on how SEM analyses are undertaken, primarily focusing on the few aspects of conducting SEM that differ from CFA.

Model specification. SEM requires specification of a model (or models) to be

tested. As in CFA, this is done by specifying which relations between variables are fixed at some value (often zero) or left free to vary (to be estimated during model fitting). SEM models have two components: the measurement model and the structural model. The measurement model refers to the pattern of relations between latent and measured variables. The structural model refers to the pattern of relations among latent variables. The inclusion of endogenous latent variables (i.e., latent DVs that receive causal influences) in SEM adds some features that were not present in CFA models. As shown in [Figure 19.8](#), a new type of error term (latent variable “edv”) is added. This term reflects residual variance in the latent variable not accounted for by the other latent variables hypothesized to influence it in the model (i.e., errors in equations). Also, the scale of the endogenous latent variable must be set, either by fixing its variance at 1.0 (as in [Figure 19.8](#)) or by fixing at 1.0 a path to one of the latent variable's indicators.

Many of the same issues that arise in model specification in CFA also arise in SEM. The model must be identified. Couching the research question in terms of comparisons among a priori alternative models can also be beneficial. In SEM, researchers sometimes have substantive interest in testing hypotheses about the number and nature of common factors underlying the data (i.e., the measurement model), but in many cases, these questions have already been thoroughly explored or are not of particular interest. Instead, the primary goal is to understand the relations among the latent variables (i.e., the structural model).⁹ Most commonly, alternative models in SEM focus on the structural model. In some cases, alternatives might include additional paths or might omit paths among latent variables. In other cases, alternatives might differ in the direction of one or more paths. Researchers also sometimes specify alternative models that involve setting constraints on specific parameter estimates of a model. As in CFA, “constrained” models are nested within “unconstrained” models, thus permitting statistical tests of differences in fit (e.g., using the likelihood ratio statistic). This allows the researcher to test specific hypotheses regarding parameter estimates in a given model.

Model fitting, evaluation, and modification. The process for model fitting is identical to that of CFA. The same indices of fit can be used, and the same general issues arise. Likewise, the criteria used in CFA model evaluation (model fit, interpretability of parameter estimates, and parsimony of the model) are also typically used in SEM. A researcher can separately assess fit of the measurement and structural portions of the model (Anderson & Gerbing, [1988](#)). At a minimum, researchers should realize that overall model fit in SEM is a joint function of the plausibility of both aspects of the model (Schumacker & Lomax,

2010).

One difficulty that sometimes arises in SEM (but not typically in CFA) is the problem of equivalent models (for a detailed discussion of this issue, see Fabrigar & Wegener, 2009; MacCallum et al., 1993). For some models there exist one or more mathematically equivalent models that cannot be distinguished on the basis of their fit to the data. No comprehensive set of rules has been developed for determining when an alternative model will be equivalent to a preferred model. However, rules for generating classes of equivalent models have been developed. These rules allow a researcher to determine changes in the direction of relations among latent variables that will result in a mathematically equivalent model. Because the resulting equivalent models will always have the same number of free parameters as the original model, they will have not only identical goodness of fit but also identical statistical parsimony. Thus, when such equivalent models exist, researchers will need to provide a rationale for preferring one model based on interpretability of parameter estimates or conceptual plausibility (for examples, see Fabrigar & Wegener, 2009; MacCallum et al., 1993).

The issues arising in modification of CFA models are equally applicable to SEM models. If anything, the temptation to modify models is greater in SEM, because both the measurement and structural portions of the model provide opportunities for modification. Unfortunately, the problems inherent in empirically driven model modifications in CFA are also true of SEM.

Types of Hypotheses

SEM provides an extremely flexible mathematical framework in which a wide range of different causal relations can be represented and tested. Virtually all of the types of hypotheses discussed thus far can be examined in SEM. Furthermore, these hypotheses can be tested when the constructs are represented by measured variables (i.e., when there are only single indicators assessing each construct) or when constructs are represented by latent variables.

Direct cause. SEM affords the ability to test virtually any form of direct causal relation (a single IV affecting a single DV, multiple IVs influencing a single DV, a single IV influencing multiple DVs, or multiple IVs affecting multiple DVs). Furthermore, the researcher can specify all IVs as correlated, all IVs as uncorrelated, or some subset of IVs as correlated. SEM also allows one to specify models in which a given IV only influences a subset of DVs.

Another critical (and defining) feature of SEM is the ability to specify IVs and/or DVs as latent variables. If the researcher has two or more measured variables that are designed to assess the same latent variable, the researcher can specify causal and/or noncausal relations between this latent variable and other variables in the model. A single measured variable that reflects a given construct can be represented in one of two ways. Some SEM programs represent such constructs as if they were latent variables, but with only the single measure as a perfect (errorless) indicator of the latent variable. Other programs do not require specification of a latent variable, including measured variables themselves as part of the structural model. Models specified in either manner are mathematically equivalent, and a latent variable with a single indicator is not a latent variable in any real sense, because there is no representation of error.

Relative strength of direct causes. It is very easy to compute tests of the relative strength of direct causes in SEM if latent variables have been set on a common scale of measurement. For example, a researcher could readily test a hypothesis regarding relative impact of the two IVs on the DV in [Figure 19.8](#). This could be done by testing the difference in the likelihood ratio test statistic between a model with these parameter estimates constrained to be equal and a model without this constraint. A significant difference in the likelihood ratio test statistics of these models indicates the coefficients are not the same. Comparisons of the relative impact of an IV on two DVs can also be tested using the same approach. Of course, just as in the context of regression, anytime a difference is observed in the relative strength of two direct causal effects, such a difference is potentially open to multiple interpretations.

Mediation. One of the most common and powerful uses of SEM is exploration of mediation. For example, consider a researcher who wants to assess mediation of frustration effects on aggression by examining not only frustration effects on negative emotions (one possible mediator), but also effects of negative emotions on interpretation of actions by others (with different interpretations leading to different levels of ultimate aggression). This multistep mediation would become unwieldy for traditional regression-based tests (but see Hayes et al., [2011](#) for bootstrapping approaches to testing such models in regression). However, SEM easily handles such a model. In fact, virtually any mediational pattern among measured variables or latent variables can be specified in SEM (a single IV influencing a single mediator, which in turn influences a single DV, multiple IVs, multiple mediators, multiple DVs, or any combination of these). As with direct causal SEM models, the researcher can

allow for correlations among IVs and can specify a model in which a specific IV influences only a subset of mediators and/or a given mediator influences only a subset of DVs.

There are several important differences between SEM and other statistical procedures used to test mediation (Judd et al., Chapter 25 in this volume). SEM allows more flexibility in specifying different mediational models. Also, because SEM fits a single model to the data rather than requiring a series of analyses, it is possible to derive indices of model fit for the entire model. Finally, SEM allows researchers to specify latent variables and causal relations among them.

Moderation. Moderational hypotheses have typically been tested in SEM in one of two ways. One method is through the use of interaction terms. When all variables in a model are single measured variables presumed to be perfect indicators of the constructs of interest, moderation can be tested using procedures analogous to those used in multiple regression. That is, a model can then be specified that includes the IVs and the product of the IVs as causal influences on the DV(s) (along with correlations among the IV terms). Such models are mathematically equivalent to regression models including interaction terms.

It is also possible to specify SEM models with interactions among actual latent variables. There are potential advantages of specifying latent variable interaction models over measured variable interaction models. Because product variables are less reliable than are the component variables used to form them, it is often more difficult to detect moderator effects than direct causal effects (McClelland & Judd, 1993). Because latent variables do not include random error, interaction terms created from such variables have the potential to be more sensitive than measured-variable interaction terms.

A number of approaches have been suggested for specifying and testing such models (e.g., Kenny & Judd, 1984; Marsh et al., 2007; Ping, 1996). Detailed discussion of these approaches is beyond the scope of this chapter (but see Judd et al., Chapter 25 in this volume). However, a central feature of these approaches is the creation of a latent variable interaction term whose indicators are product terms of all possible combinations of the indicators of the two latent variables hypothesized to interact with one another. This approach to testing moderation is conceptually sensible but can prove difficult to implement with complex interaction models, because the number of parameters associated with the interaction latent variables is likely to become very large.

A considerably simpler approach, especially if a moderator is dichotomous, is to conduct a multisample SEM analysis (much like a multisample CFA analysis; see Widaman & Grimm, Chapter 20 in this volume). This approach involves no computation of interaction variables and is analogous to the conduct of separate regression analyses at each level of a moderator. In this approach, a researcher categorizes participants according to the moderator. A covariance matrix among the measured variables other than the moderator variable is computed for each group. Then, a model specifying the IV(s) influencing the DV(s) is simultaneously fit to each of the matrices, and the parameter values for the model obtained from the different matrices are compared.

As discussed by Widaman and Grimm (Chapter 20 in this volume), it is not meaningful to make comparisons of parameters in the structural model across groups unless equivalency of measurement model across groups is first established. Assuming factorial invariance is established, the direct causal model of interest is specified and simultaneously fit to each group, allowing structural parameter estimates to vary across groups (the unconstrained model). Variation in the impact of the IV on the DV is tested by comparing the unconstrained model with a constrained model in which the path between the IV and the DV must be equivalent across groups. A significant difference in fit between the constrained and unconstrained models indicates that a moderation effect is present.

Multisample analyses provide a relatively simple method for testing moderational hypotheses in SEM. However, there are limitations. First, such analyses necessarily involve treating the moderator variable as categorical even if it is measured on a continuous scale. This limitation exists because there will generally not be sufficient sample size to define a group for each score of the moderator variable. Additionally, even if the sample size is large enough, simultaneously fitting a model to many groups is extremely unwieldy and might pose difficulties to parameter estimation. A second limitation is that there is no way to represent the moderator variable as a latent variable in multisample analyses (because categorization of individuals on the moderator variable must be based on observed scores). Thus, some of the benefits of testing moderational hypotheses in SEM are lost (although it is still possible to treat other IVs and DVs as latent variables). Finally, multisample approaches become rather unwieldy for testing more complex moderational hypotheses (e.g., three-way interactions) because the number of groups becomes large and no direct overall test exists for more than a single moderator.

Hybrid hypotheses. Hybrid hypotheses can also be tested in SEM. The simplest way of addressing moderated mediation questions would be through multisample analysis strategies. That is, the mediational model of interest is specified and simultaneously fit to covariance matrices for different groups (determined by scores on a moderator). Tests between groups on the paths in the structural model can then be conducted by including constraints in the model. SEM could also implement regression-like tests of moderated mediation and mediated moderation when single rather than multiple indicators of constructs are used.

A final class of hypotheses that has not been previously discussed is moderation of noncausal relations among constructs of interest. In some cases, a researcher might be interested in examining differences in the factor loadings of a model across different groups, differences in the unique variances across groups, or differences in the correlations among factors across groups. Such hypotheses can be tested by multisample CFA (e.g., see Judd & Krosnick, 1982).

Design Issues in the Use of SEM

Consideration of study design can greatly enhance the utility of results obtained from SEM. Many of the design issues in CFA, such as selection of measured variables and the nature and size of the sample, are equally relevant to SEM. However, because SEM generally involves formulating models that postulate certain causal relations among latent variables, additional considerations are necessary. Simply because an SEM model postulates causal relations does not ensure that such assumptions are correct. Thus, a researcher should consider how design features might strengthen the basis for making causal inferences. Experimental manipulation of key constructs provides a strong basis for causal inferences. Also, incorporating longitudinal features into the design can sometimes assist in establishing conditions of causality, although time ordering alone is often insufficient (MacCallum et al., 1993; West et al., Chapter 4 in this volume).

SEM has proven to be one of the most important developments in quantitative methodology over the past 40 years. SEM can be used to specify a wide range of models involving direct causal or mediational relations among measured variables and latent variables. Many of these models would be difficult and sometimes impossible to specify in other common methods of analysis such as multiple regression and ANOVA.

Summary of Methods Addressing Causal Hypotheses

A variety of approaches can be brought to bear on hypotheses of cause in nonexperimental data. The techniques most commonly used (multiple regression, MLM, SEM) afford researchers some leverage in making a case for causal impact of the hypothesized IV(s). Regression utilizes a rather simple mathematical model, but complex interaction effects fit easily within the framework, and it can be adapted to various mediational and moderational questions. Moreover, regression analyses can often be usefully applied with relatively small samples (at least substantially smaller than typically recommended for more complex analyses, such as SEM). Related to the conditions of isolation, association, and direction, the primary benefit of multiple regression is an increase in ability to isolate effects of the proposed IV on the DV. By measuring and statistically controlling alternative influences on the DV, confidence may increase that the influence of an IV does not stem from other constructs included in the model. Overall confidence in the causal impact of the IV increases to the extent that all plausible alternative causes are included in the model. Because association assumes relation between the IV and the DV separate from alternative influences on the DV, increasing isolation also aids in meeting the condition of association. Especially compared to techniques limited to categorical IVs (e.g., ANOVA), the ability of regression to utilize continuous measures of IVs increases the ability to find associations between IV and DV.

However, regression also has limitations in this regard. Measures are treated as perfect (errorless) indicators of the constructs under study, even though there is almost always some error in psychological measures. Such errors underestimate relations between constructs and thus limit the extent to which complete partialing can take place. In addition to (lack of) representation of error in the model, regression techniques can be cumbersome for many kinds of advanced research questions. Even simple mediation questions require several regressions, and this gets more complex as the number of mediators increases (although macros for bootstrapping approaches to test such questions have greatly simplified the matter when using regression).

The primary benefits of MLM for addressing issues of causation are largely shared with the traditional regression technique. Of course, to the extent that the multilevel approach improves estimation and testing of certain effects (because of error terms for lower-level slopes and intercepts; see Schoemann et al., Chapter 21 in this volume), they are better able to isolate effects and establish association. The utility of regression and MLM rests on the ability of researchers

to include in the model alternative influences on the DV that can be statistically controlled. Of course, such techniques are only successful to the extent that alternative causes of variation in the DV are exhaustively represented and measured well. Statistical techniques will be unable to isolate hypothesized influences on the DV when poor measures are utilized or alternative plausible causes are omitted (Wegener, Downing, Krosnick, & Petty, 1995).

The SEM latent-variable approach facilitates isolating causes and finding associations between IVs and DVs by representing error of measurement in the model. When measures contain random error, relations among IVs are underestimated. Because latent variables are free of random error, SEM affords better statistical control of other IVs in the model, thereby allowing a researcher to establish the condition of isolation more effectively. The ability to establish association can also be improved through the use of latent variables. In addition to leading to underestimation of relations among IVs, random error also attenuates relations between IVs and DVs. Thus, random error can mask true relations between IVs and DVs. The analysis of latent variables can help alleviate this problem. Inclusion of random error of measurement in the model is also more realistic than assuming an absence of error.

Most common statistical procedures (e.g., regression) are used for parameter estimation (e.g., to assess the impact of an IV on a DV) and are based on mathematical models that are nonfalsifiable (i.e., for which fit of the overall model cannot be assessed). Although parameter estimation is extremely useful, in some contexts a researcher might also be interested in assessing the plausibility of a hypothesized pattern of relations among constructs. SEM allows a researcher to specify falsifiable models and provides a wide variety of model fit indices to address such questions. Model fit can also be useful when a researcher wishes to compare two models that postulate different directional relations among constructs. Thus, SEM affords some advantage in assessing the condition of directionality (although the existence of equivalent or near-equivalent models will sometimes limit this potential advantage). As noted repeatedly in this volume, design features of the study (e.g., manipulation of crucial hypothesized causal factors) often have the most direct impact on the causal inferences that researchers can make.

Conclusions

Numerous statistical procedures can be used to test noncausal and causal

hypotheses. These procedures are not exclusive to either experimental or nonexperimental studies. Rather, these procedures are best conceptualized in terms of the sorts of hypotheses they can be used to test. Moreover, the strength with which hypotheses can be tested with these various methods is more a function of the design of the study than the statistical procedure being used. Even so, each statistical method has its own strengths and limitations, as well as particular issues to which researchers must attend. In many cases, clear communication of the analyses and results requires specification of the details of the analysis. In part, we believe that greater attention to the methodological literature on best practices would be beneficial, and we have tried to point the interested reader to relevant methodological discussions.

References

- Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103, 411–423.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, 11, 142–163.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bollen, K. A. (2007). Interpretational confounding is due to misspecification, not to type of indicator: Comment on Howell, Breivik, and Wilcox (2007). *Psychological Methods*, 12, 219–228.
- Borg, I., & Groenen, P. (2005). *Modern multidimensional scaling* (2nd ed.). New York: Springer-Verlag.
- Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research*, 38, 25–56.
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, 21, 230–258.

- Browne, M. W., Cudeck, R., Tateneni, K., & Mels, G. (2010). *CEFA: Comprehensive exploratory factor analysis*. Version 3.04 [Computer software and manual]. Retrieved from <http://faculty.psy.ohio-state.edu/browne/> on August 26, 2013.
- Browne, M. W., MacCallum, R. C., Kim, C., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, 7, 403–421.
- Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, 27, 509–540.
- Carroll, J. D., & Wish, M. (1974). Models and methods for three-way multidimensional scaling. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology: Measurement, psychophysics, and neural information processing* (Vol. 2, pp. 57–105). San Francisco: W. H. Freeman.
- Clark, J. K., Wegener, D. T., & Fabrigar, L. R. (2008). Attitudinal ambivalence and message-based persuasion: Motivated processing of proattitudinal information and avoidance of counterattitudinal information. *Personality and Social Psychology Bulletin*, 34, 565–577.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Crites, Jr., S. L., Fabrigar, L. R., & Petty, R. E. (1994). Measuring the affective and cognitive properties of attitudes: Conceptual and methodological issues. *Personality and Social Psychology Bulletin*, 20, 619–634.
- Cudeck, R., & O'Dell, L. L. (1994). Applications of standard error estimates in unrestricted factor analysis: Significance tests for factor loadings and correlations. *Psychological Bulletin*, 115, 475–187.
- DeSteno, D., Petty, R. E., Wegener, D. T., & Rucker, D. D. (2000). Beyond valence in the perception of likelihood: The role of emotion specificity. *Journal of Personality and Social Psychology*, 78, 397–416.
- DeSteno, D. A., & Salovey, P. (1997). The effects of mood on the structure of

- the self-concept. *Cognition and Emotion*, 11, 351–372.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Chichester, UK: John Wiley.
- Fabrigar, L. R., & Wegener, D. T. (2009). Structural equation modeling. In J. P. Stevens (Ed.), *Applied multivariate statistics for the social sciences* (5th ed., pp. 537–582). New York: Routledge.
- Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. New York: Oxford University Press.
- Fabrigar, L. R., Wegener, D. T., MacCallum R. C., & Strahan, E. J. (1999). Evaluating the use of factor analysis in psychological research. *Psychological Methods*, 4, 272–299.
- Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75–109). San Diego, CA: Academic Press.
- Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology*, 69, 153–166.
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Greenacre, M. (2007). *Correspondence analysis in practice* (2nd ed.). Boca Raton, FL: Chapman & Hall.
- Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number-of-factors rules with simulated data. *Multivariate Behavioral Research*, 17, 193–219.
- Halberstadt, J. B., & Niedenthal, P. M. (1997). Emotional state and the use of stimulus dimensions in judgment. *Journal of Personality and Social Psychology*, 72, 1017–1033.
- Harshman, R. A. (1984). “How can I know if it's real?” A catalog of diagnostics for use with three-mode factor analysis and multidimensional scaling. In H. G. Law, C. W. Snyder, Jr., J. A. Hattie, & R. P. McDonald (Eds.), *Research*

methods for multimode data analysis (pp. 566--571). New York: Praeger.

- Hayes, A. F., Preacher, K. J., & Myers, T. A. (2011). Mediation and the estimation of indirect effects in political communication research. In E. P. Bucy, & R. Lance Holbert (Eds.), *Sourcebook for political communication research: Methods, measures, and analytical techniques* (pp. 434-- 465). New York: Routledge.
- Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, 12, 205–218.
- Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424–153.
- Humphreys, L. G., & Montanelli, Jr., R. G. (1975). A investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, 10, 193–205.
- John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality* (2nd ed., pp. 102–138). New York: Guilford Press.
- Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions*. New York: Cambridge University Press.
- Judd, C. M., & Krosnick, J. A. (1982). Attitude centrality, organization, and measurement. *Journal of Personality and Social Psychology*, 42, 436–447.
- Judd, C. M., & McClelland, G. H. (1998). Measurement. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 180–232). New York: McGraw-Hill.
- Judd, C. M., McClelland, G. H., & Ryan, C. S. (2008). *Data analysis: A model comparison approach* (2nd ed.). New York: Routledge.
- Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, 96, 201–210.
- Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 233–265). New York: McGraw-Hill.

- Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1–27.
- Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Newbury Park, CA: Sage.
- MacCallum, R. C. (1974). Relations between factor analysis and multidimensional scaling. *Psychological Bulletin*, 81, 505–516.
- MacCallum, R. C. (1977). Effects of conditionality on INDSCAL and ALSCAL weights. *Psychometrika*, 42, 297–305.
- MacCallum, R. C. (1988). Multidimensional scaling. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 421–145). New York: Plenum Press.
- MacCallum, R. C. (1990). The need for alternative measures of fit in covariance structure modeling. *Multivariate Behavioral Research*, 25, 157–162.
- MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, 11, 19–35.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130–149.
- MacCallum, R. C., & Hong, S. (1997). Power analysis in covariance structure modeling using GFI and AGFI. *Multivariate Behavioral Research*, 32, 193–210.
- MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, 32, 215–253.
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111, 490–504.
- MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, 114, 185–199.
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size

in factor analysis. *Psychological Methods*, 4, 84–99.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Taylor & Francis.

Marsh, H. W., Wen, Z., Hau, K., Little, T. D., Bovaird, J. A., & Widaman, K. F. (2007). Unconstrained structural equation models of latent interactions: Contrasting residual-and mean-centered approaches. *Structural Equation Modeling*, 14, 570–580.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114, 376–390.

Muller, D., Judd, C. M., Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, 89, 852–863.

Muthén, B. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 204–234). Newbury Park, CA: Sage.

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, 41, 847–855.

Petty, R. E., Cacioppo, J. T., & Schumann, D. (1983). Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of Consumer Research*, 10, 135–146.

Petty, R. E., Wegener, D. T., Fabrigar, L. R., Priester, J. R., & Cacioppo, J. T. (1993). Conceptual and methodological issues in the Elaboration Likelihood Model of persuasion: A reply to the Michigan State critics. *Communication Theory*, 3, 336–363.

Ping, R. A. (1996). Latent variable interaction and quadratic effect estimation: A two-step technique using structural equation analysis. *Psychological Bulletin*, 119, 166–175.

Preacher, K. J. (2011). Multilevel SEM strategies for evaluating mediation in three-level data. *Multivariate Behavioral Research*, 46, 691–731.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods*,

Instruments, and Computers, 36, 717–731.

- Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40, 879–891.
- Preacher, K. J., & MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior Genetics*, 32, 153–161.
- Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Assessing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42, 185–227.
- Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling*, 18, 161–182.
- Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15, 209–233.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.
- Rosenberg, M. J., & Hovland, C. I. (1960). Cognitive, affective, and behavioral components of attitudes. In C. I. Hovland & M. J. Rosenberg (Eds.), *Attitude organization and change: An analysis of consistency among attitude components* (pp. 1–14). New Haven, CT: Yale University Press.
- Rusbult, C. E., Onizuka, R. K., & Lipkus, I. (1993). What do we really want?: Mental models of ideal romantic involvement explored through multidimensional scaling. *Journal of Experimental Social Psychology*, 29, 493–527.
- Rusbult, C. E., & Zembrodt, I. M. (1983). Responses to dissatisfaction in romantic involvements: A multidimensional scaling analysis. *Journal of Experimental Social Psychology*, 19, 274–293.
- Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 181–204). Newbury Park, CA: Sage.

- Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). *Introduction to multidimensional scaling: Theory, methods, and applications*. New York: Academic Press.
- Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). Hillsdale, NJ: Erlbaum.
- Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, 7, 422–445.
- Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, 89, 845–851.
- Takane, Y., Young, F. W., & De Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least-squares method with optimal scaling features. *Psychometrika*, 46, 389–405.
- Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, 34, 421–459.
- Tucker, L. R., & MacCallum, R. C. (1997). *Exploratory factor analysis*. Unpublished manuscript.
- Wegener, D. T., Clark, J. K., & Petty, R. E. (2006). Not all stereotyping is created equal: Differential consequences of thoughtful and nonthoughtful stereotyping. *Journal of Personality and Social Psychology*, 90, 42–59.
- Wegener, D. T., Downing, J., Krosnick, J. A., & Petty, R. E. (1995). Strength-related properties of attitudes: Measures, manipulations, and future directions. In R. E. Petty and J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 455–187). Mahwah, NJ: Erlbaum.
- Wegener, D. T., & Fabrigar, L. R. (2000). Analysis and design for nonexperimental data: Addressing causal and noncausal hypotheses. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 412–450). New York: Cambridge University Press.
- Wegener, D. T., Petty, R. E., & Smith, S. M. (1995). Positive mood can increase

or decrease message scrutiny: The hedonic contingency view of mood and message processing. *Journal of Personality and Social Psychology*, 69, 5–15.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 56–75). Newbury Park, CA: Sage.

Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, 28, 263–311.

Wish, M., & Carroll, J. D. (1974). Applications from individual differences scaling to studies of human perception and judgment. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2, pp. 449–491). New York: Academic Press.

Zhang, G., Preacher, K. J., & Luo, S. (2010). Bootstrap confidence intervals for ordinary least squares factor loadings and correlations in exploratory factor analysis. *Multivariate Behavioral Research*, 45, 104–134.

Zhong, X., & Yuan, K. (2011). Bias and efficiency in structural equation modeling: Maximum likelihood versus robust methods. *Multivariate Behavioral Research*, 46, 229–265.

¹ One term often used instead of nonexperimental data is “correlational data.” The former seems much more accurate because it does not use a type of analysis to describe a characteristic of data. Correlational analysis is equally applicable to data collected using experimental or nonexperimental methods.

² A superficially similar model that has created much confusion is the principal component model. Researchers using principal components analysis (PCA) often assume it is simply a form of EFA that produces similar results. In point of fact, PCA is based on a different underlying mathematical model, was originally designed for somewhat different goals, and in some cases can produce substantively different results (see Fabrigar et al., 1999; Widaman, 1993). PCA is an appropriate data-reduction procedure when a researcher simply wishes to reduce a larger number of variables to a smaller set of composite variables that

retain as much information from the original variables as possible. However, EFA is more appropriate for identifying latent psychological constructs. Researchers must be careful, because common statistical software (such as SAS and SPSS) includes PCA as the default setting in their EFA routines. Thus many researchers have likely conducted PCA despite thinking that they were conducting EFA (for more detailed discussion, see Fabrigar & Wegener, 2012).

³ To fit the model, constraints are placed on the EFA at the factor extraction phase (i.e., when a second common factor is included, one of the measured variables is set with a loading of zero on the second factor; when a third common factor is included, the same measure and one other are set to load at zero on the third factor, etc.). When EFA solutions are rotated, however, factor loadings for all measured variables can take on nonzero values (see later discussion of rotation).

⁴ Methods for calculating standard errors, confidence intervals, and statistical tests of parameter estimates have also been developed for Ordinary Least Squared (OLS) procedures. Some such tests share the ML assumption of multivariate normality of measures (Browne, Cudeck, Tateneni, & Mels, 2010). However, bootstrapping methods have also been developed to generate confidence intervals for factor loadings and interfactor correlations in OLS EFA (Zhang, Preacher, & Luo, 2010).

⁵ Each of the preceding discussions, with the exception of use of model fit in determining the number of factors, is also applicable to the use of PCA. However, PCA also often provides poorer simple structure than EFA does, and can substantially underestimate the correlations among factors when oblique rotations are used (e.g., see Widaman, 1993; Fabrigar et al., 1999).

⁶ We confine our discussion to model-fitting procedures that assume at least interval level data. Procedures for fitting CFA (and covariance structure models more generally) to data with ordinal or nominal properties have also been developed (e.g., Muthen, 1993).

⁷ When examining moderator relations, there is often ambiguity with respect to which IV might be designated as the moderator. In some cases, there might be

conceptual reasons that make it useful to treat one IV as moderating the influence of the other IV on the DV. In other situations, it might be equally meaningful to designate either of the IVs as the moderator. When testing moderator hypotheses using interaction terms, this decision has little practical consequence, as the interaction test does not require specifying which IV is the moderator. However, when other procedures are used (e.g., dividing participants into groups based on their scores on the moderator variable and then testing the impact of the IV on the DV within each group), this decision can have consequences for the manner in which the analysis is conducted.

⁸ In multiple regression, the overall R^2 of the model is sometimes conceptualized as an index of model fit. In the strict sense of how model fit is typically defined in latent variable modeling (i.e., the extent to which the model accounts for the covariances among a set of measured variables), R^2 does not reflect model fit. Rather it constitutes a parameter estimate in the model. It is an alternative expression of the residual variance in the DV not accounted for by the predictor variables. The regression model cannot distinguish whether residual variance in the DV results from the failure of the underlying constructs represented by the IVs to account for variance in the underlying construct represented by the DV (error in equations) versus a failure of the IVs and DVs to effectively represent the underlying constructs (error in variables). Thus, R^2 does not reflect the plausibility of the underlying causal assumptions of the model. Likewise, R^2 does not reflect any aspect of the relations among the IVs, which are part of the overall model and thus relevant to the overall fit of the model.

⁹ SEM models typically conceptualize latent variables as influencing the measured variables. In such models, measured variables are referred to as effects indicators. However, for some constructs, this assumption is questionable. For example, socioeconomic status (SES) might be conceptualized in the opposite causal direction, with indicators such as income and level of education (referred to as causal indicators) combining to produce SES. There continues to be substantial debate regarding the utility of causal indicator models (e.g., see Bollen, 2007; Howell, Breivik, & Wilcox, 2007a).

Chapter twenty Advanced Psychometrics

Confirmatory Factor Analysis, Item Response Theory, and the Study of Measurement Invariance

Keith F. Widaman and Kevin J. Grimm*

Factor analytic techniques, including common factor analysis and principal component analysis, are arguably the most popular and useful methods for identifying dimensions of individual difference that can account for behavior in a given domain. More than a century ago, Spearman (1904) proposed novel calculations that became known as common factor analysis. Spearman (1904, 1927) used results of his analyses to argue for the existence of ‘*g*,’ or general intelligence, the single ability dimension common to all tests of mental ability. Later, Thurstone (1931, 1935, 1938, 1947) generalized the common factor model to include multiple factors. Then, Thurstone and his colleagues (e.g., Thurstone, 1938; Thurstone & Thurstone, 1941) identified and replicated seven dimensions of mental ability that became known as primary mental abilities. Currently, the most commonly accepted structure of the mental ability domain is often identified as the Cattell-Horn-Carroll (CHC) model, based on work by Cattell (1963, 1971), Horn (e.g., Horn & Hofer, 1992; Horn & Noll, 1997), and Carroll (1993). In the CHC model, three strata of factors are posited. At the lowest stratum reside perhaps 30 or so primary or narrow abilities, such as Verbal Comprehension and Numerical Facility. At the second stratum level, eight or nine dimensions have been replicated, including factors for Fluid Intelligence (*Gf*), Crystallized Intelligence (*Gc*), and General Speediness (*Gs*). At the third stratum, a single, overarching dimension has been posited – identified as general intelligence, or *g* – although some dispute whether this dimension can be justified theoretically (e.g., Horn & McArdle, 2007; Horn & Noll, 1997).

In the domain of personality, common factor analysis and principal component analysis have also been the primary tools for identifying a replicable set of dimensions. The most widely accepted taxonomy of dimensions in the personality domain is known as the Big 5 (or the Five Factor model), with dimensions of Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. Although one can trace the origins of the Big 5

personality constructs to studies from the mid-1940s or even earlier (Thurstone, 1934; Fiske, 1949), a consistent stream of research studies from the late 1950s to the present has offered convincing evidence that the Big 5 dimensions can be recovered in data from many sources (e.g., self-reports, observer reports, reports by family members or friends) and across an impressive number of cultures (Goldberg, 1993).

More recently, item response theory (IRT) models have been used to study dimensions of individual differences in many domains, including mental abilities and personality. Technical, theoretical work on IRT models can be traced back about 60 years, so IRT modeling has a shorter history than does factor analysis. But, the use of IRT models in psychology is growing rapidly and is the basis for much work on standardized tests in many areas of study in psychology and associated sciences.

Once dimensions of a domain of empirical phenomena have been replicated sufficiently to be accepted as a useful taxonomy of constructs for the domain, the issue of measurement invariance comes to the fore. Measurement invariance holds if a given construct is measured in the same fashion or in a comparable way across groups. This goal is usually accomplished by showing that a latent variable representing a construct is related to its manifest indicators, such as items, in the same way across groups. Measurement invariance would not hold, for example, if items for the “talkativeness” or “gregariousness” facet of Extroversion were more strongly related to the latent trait than were items for the “social potency” facet in one group, but items for “social potency” were related more strongly to the Extroversion latent trait than were “gregariousness” items in a comparison group. Obtaining an Extroversion scale score across all items and attempting to make comparisons across groups would amount to making “apples vs. oranges” comparisons, as the latent trait was measured differently across groups and therefore had a different interpretation across groups. But if the relations of “gregariousness” and “social potency” items to the Extroversion trait were the same across groups, then comparisons across groups on the latent trait would be valid, because the latent trait or construct was assessed in a comparable way in each group.

The primary focus of the current chapter is to introduce the use of confirmatory factor analysis (CFA) and IRT modeling, particularly as each is used to evaluate measurement invariance of assessment devices. First we cover CFA, which is a special case of the common factor model. In this section we discuss basic ideas with regard to model specification and evaluation and how

measurement invariance is pursued in CFA models. Then we discuss parallel issues with regard to IRT models, including both basic forms of IRT model and the study of measurement invariance within such models. In each of these sections we provide illustrations of analyses with empirical data to demonstrate how to fit such models and evaluate results. Sophisticated modeling of data is becoming ever more common in research in personality and social psychology, and we trust our efforts to elucidate measurement techniques in CFA and IRT models will provide a useful introduction for researchers to pursue methodological approaches at the cutting edge of quantitative research.

Confirmatory Factor Analysis

CFA involves, basically, fitting restricted forms of exploratory factor analysis (EFA) models to data. EFA is a method for locating and characterizing the dimensions of individual differences in a domain of behavior or content without imposing an a priori structure on the dimensions. Indeed, Thurstone (1947) argued that EFA was of greatest use in initial explorations of a domain, because a successful analysis would allow one to discover and isolate the major dimensions of individual differences within the domain. Once the primary dimensions within a domain had been identified and replicated in several studies, later research could investigate the bases of individual differences on these dimensions.

The impetus for developing EFA is a theoretical orientation that deserves mention. Both Spearman (1904) and Thurstone (1947) argued that scientific goals are furthered most by understanding the varied phenomena in a domain in terms of a smaller number of dimensions that can account for the phenomena in an economical fashion. But economy of representation was not the only driving force or even the principal driving force. Instead, common factor analysis was considered the optimal way to isolate the underlying fundamental behavioral processes that generate the myriad forms of behavior in a particular domain of content. Some authors (e.g., Goldberg & Velicer, 2006; Velicer & Jackson, 1990) have argued that principal component analysis and EFA provide equivalent representations of data and that the former method is conceptually and computationally simpler than the latter. Furthermore, some researchers, including Goldberg (1993; Goldberg & Velicer, 2006) and Costa and McCrae (e.g., McCrae & Costa, 1987) tended to use only principal component analysis in their research. However, Widaman (1993, 2007) demonstrated that principal component analysis can often lead to seriously distorted representations of data

if used as a prelude to CFA, whereas EFA representations are well coordinated with CFA representations. As a result, this chapter will concentrate on EFA and CFA procedures, and principal component analysis approaches will not be discussed further in any detail.

Most factor analytic work on the Big 5 dimensions of personality has used EFA procedures. As explained by Fabrigar and Wegener (Chapter 19 in this volume), EFA involves a number of steps, including communality estimation, deciding on the number of factors, rotating factors to an interpretable orientation, and so forth. Using EFA, researchers occasionally disagree on the propriety of certain procedures, such as orthogonal versus oblique rotation of factors. In addition, evaluating the results of EFAs can vary, as some researchers treat small but trivial factor loadings (e.g., loadings in the .10 to .25 range) as if they were zero, whereas other researchers see these small loadings as worthy of note.

CFA arose in the late 1960s with development of efficient algorithms for estimating parameters in factor analysis models and recognition that certain restrictions, such as fixing certain factor loadings to be precisely zero, were hypotheses that could be tested statistically. Rather than arguing whether a loading of .15 was or was not essentially zero, one could fix the loading precisely at zero and evaluate whether this restriction worsened model fit significantly. The first general presentation of CFA was published by Jöreskog (1969), who discussed how a CFA model places restrictions on the EFA representation of data. These restrictions can be based on prior research or theory, but must be invoked a priori to enable simple interpretation of measures of model fit to the data. Early contributions to CFA by Jöreskog (1969, 1971a, 1971b) were quickly followed by a spate of publications by researchers who understood the power of confirmatory factor analysis to provide statistical tests of hypothesized factor structures.

CFA Model

Data model. The fundamental equation of common factor analysis, both EFA and CFA, is a data model that specifies the relations of the underlying factors or latent variables (LVs) to the observed or manifest variables (MVs) in an analysis. This model represents each MV $Y_j (j = 1, \dots, p)$ as an additive, linear function of one or more of the k underlying, unobserved common LVs $Z_{c_{ki}} (k = 1, \dots, r)$ and the unique LV for the MV $Z_{u_{ji}}$, or

$$Y_{ji} = \tau_j + \lambda_{j1} Z_{c_{1i}} + \lambda_{j2} Z_{c_{2i}} + \cdots + \lambda_{jr} Z_{c_{ri}} + \lambda_{ju} Z_{u_{ji}} \quad (20.1)$$

where Y_{ji} is the score of person i ($i = 1, \dots, N$) on MV j , τ_j is the intercept (or mean) of MV j , λ_{jk} is the loading of MV j on LV k , $Z_{c_{ki}}$ is the score of person i on common LV k , λ_{ju} is the loading of MV j on the j th unique factor, and $Z_{u_{ji}}$ is the score of person i on the j th unique factor. Note that in Equation 20.1, the unique factor functions as a residual term, representing variation in Y_{ji} remaining after the common LVs explain variance in the manifest variable. Some authors identify the unique factors as error terms, which implies that these terms reflect random error alone; we prefer the term “unique factors” because these residual terms each consist of a sum of effects of random error and systematic sources of variance unrelated to the common LVs.

The CFA data model thus embodies the mathematical representation of p MVs in terms of, or as linear functions of, r common factors (with $r < p$) and p unique factors. The r common factors are identified as common factors because they have effects on more than a single MV. Thus, these factors represent sources of individual difference that have influences *in common* on two or more MVs. The variance in MV j that is explained mathematically by the common factors is termed the communality of variable j , which is typically represented as h^2_j .

In contrast, each of the p unique factors has an effect on a single MV, and each of the p MVs has its own unique factor. The variance of the unique factor for MV j is often represented as u^2_j . The unique variance for variable j should be understood as the sum of two sources of variance: (a) specific variance, s^2_j , which is reliable variance in MV j that is linearly unrelated to the common factors; and (b) error variance, e^2_j , which is random measurement error.

Given the preceding partitioning and assuming that variance of MV j is represented as σ_j^2 , the communality of MV j is h^2_j , unique variance is u^2_j , and $h^2_j + u_j^2 = \sigma_j^2$. But reliable variance of MV j , or r_{jj} , can be represented as $r_{jj} = (h^2_j + s^2_j)/\sigma_j^2$. Thus, the communality of MV j is a lower bound estimate of its reliability. Also, the CFA model can be seen as a generalization of the classical test theory decomposition of a MV into true and error components, a generalization that contains more than one true (i.e., common factor) score and a

specific factor for each MV.

Mean and covariance structure model. Equation 20.1 can be expressed in matrix form as

$$\mathbf{Y} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\mathbf{Z}_c + \boldsymbol{\Lambda}_u\mathbf{Z}_u, \quad (20.2)$$

where \mathbf{Y} is a $(p \times 1)$ vector of MV scores for a random person, $\boldsymbol{\tau}$ is a $(p \times 1)$ vector of intercepts, $\boldsymbol{\Lambda}$ is a $(p \times r)$ matrix of loadings of the p MVs on the r common factors, \mathbf{Z}_c is an $(r \times 1)$ random vector of common factor scores, $\boldsymbol{\Lambda}_u$ is a $(p \times p)$ diagonal matrix of loadings of the p MVs on the p unique factors, and \mathbf{Z}_u is a $(p \times 1)$ random vector of unique factor scores. We typically assume that unique factor scores have expected values (i.e., population means) of zero and that unique factors are mutually uncorrelated and uncorrelated with common factors. With these assumptions, taking the expectation of Equation 20.2 leads to:

$$E(\mathbf{Y}) = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\alpha} \quad (20.3)$$

where $\boldsymbol{\alpha}$ is an $(r \times 1)$ vector of common factor means, and other symbols were defined above. The mean structure model in Equation 20.3 shows that MV means are a function of intercepts in $\boldsymbol{\tau}$ and weighted LV means in $\boldsymbol{\alpha}$. Then, taking the expectation of covariances among the MVs in Equation 20.2 results in:

$$E(\mathbf{Y}\mathbf{Y}') = \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \quad (20.4)$$

where $\boldsymbol{\Sigma}$ is the $(p \times p)$ population covariance matrix among the MVs, $\boldsymbol{\Psi}$ is the $(r \times r)$ matrix of covariances among common factors (i.e., LVs), $\boldsymbol{\Theta}$ is the $(p \times p)$ matrix, usually diagonal, of covariances among unique factors, and other symbols were defined above. Equation 20.4 is the CFA covariance structure model and represents covariances among MVs as a function of common effects of the LVs in \mathbf{Z}_c in Equation 20.2.

Obtaining a sample of N observations and computing the sample means, $\bar{\mathbf{Y}}$, and matrix of covariances among MVs, \mathbf{S} , enables one to estimate parameters in Equation 20.3, as:

$$\bar{\mathbf{Y}} \approx \hat{\boldsymbol{\tau}} + \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\mu}} \quad (20.5)$$

where carets above matrices indicate the presence of parameter estimates in these matrices, $\boldsymbol{\mu}$ is the vector of population means on MVs, and other symbols were defined above. Equation 20.5 shows that the sample means are approximated by an additive function of the MV intercepts and the product of factor loading and factor means, giving an estimate of population means on the MVs. Estimating parameters in Equation 20.4 leads to:

$$\mathbf{S} \approx \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Theta}} = \hat{\boldsymbol{\Sigma}}, \quad (20.6)$$

where carets above matrices indicate the presence of parameter estimates in these matrices, and all symbols were defined above. As shown in Equation 20.6, the sample covariance matrix is approximated by an r -dimensional CFA solution, which in turn yields an estimate of the population covariances among MVs under the assumption that the model holds in the population.

Data Requirements for CFA

Before launching analyses using CFA, data must be obtained. At a minimum, researchers should make optimal decisions about both the observations on which an analysis is based and the manifest variables to be analyzed. As an example of selection of observations, Clark and Watson and their colleagues (Clark & Watson, 1991; Watson, Clark, Weber, Assenheimer, Strauss, & McCormick, 1995; Watson, Weber, Assenheimer, Clark, Strauss, & McCormick, 1995) developed a new measure to assess anxiety and depression in both normal and clinical participants. As a result, they evaluated statistical models in samples drawn from both nonclinical and clinical populations, to ensure adequate precision of measurement in the different populations. With regard to manifest variables, Bechtoldt (1961, 1974) was one of the first to use modern CFA methods to evaluate factorial invariance across groups. He based his analyses on data from a classic study by Thurstone and Thurstone (1941), which had identified key manifest variables for important dimensions of mental ability. CFA is an ideal technique for confirming the dimensional structure of an instrument based on prior research or based on theory, for evaluating whether measurement invariance holds across groups, and for investigating group differences on latent variables. Clearer answers to such questions can be

obtained if careful decisions, along the lines of those cited earlier, are made regarding selection of participants and of the manifest variables that will be the basis of analyses.

Selection of observations. Selection of observations (i.e., selecting a sample of persons) is an underemphasized aspect of research using CFA. Selection of observations involves at least two tasks: (1) the selection mechanism, or how observations are selected from the population for inclusion in a study; and (2) the number of observations to be included in the analysis.

In factor analytic studies, treatments are typically not evaluated, so random assignment of persons to treatments is not a major issue. Instead, selection of observations from a population is a far more important matter. Differing points of view have been offered regarding selection of observations (i.e., persons). For example, Guilford (1964) argued that samples of participants should be as homogeneous as possible, so that the influence of selection variables would be minimized. In contrast, Gorsuch (1988) opined that a researcher should attempt to obtain samples of participants who vary as much as possible on the dimensions that one expected to find in a factor analysis. A middle ground position may be stated in the following way: (a) define clearly the population from which observations (e.g., participants) are to be selected; (b) note clearly whether observations are randomly and representatively sampled from the population or whether some form of nonrandom selection will occur (e.g., a sample of convenience may be unrepresentative of the population on key dimensions); and (c) describe whether any participants either decline to participate or fail to complete the entire set of tests and will be excluded from analyses for these reasons, as this may lead to unrepresentativeness of the final sample.

The second aspect of selection of observations – determining the minimum number of observations to include in an analysis – has had a varied history. In early factor analytic studies (e.g., Spearman, 1904), little or no attention was paid to the number of observations. Later, experts in factor analysis often recommended that researchers follow a particular rule – such as obtaining at least 5 times the number of observations as MVs to be analyzed or having a sample size of at least 200 observations. However, MacCallum and colleagues (MacCallum, Widaman, Preacher, & Hong, 2001; MacCallum, Widaman, Zhang, & Hong, 1999) showed that simple rules of thumb for the number of observations could not be justified if the intent of a study was to recover the population factor structure. Instead, recovery of population factors was a

complex function of the communality of variables, the number of indicators per factor, and sample size. With MVs having high communality and many indicators per factor, population factor loadings were recovered very well with sample sizes of 60 to 100 participants. In contrast, if MVs had low communalities and a small number of indicators per factor were used, samples of 400 or larger were often required to recover population loadings well. In truly exploratory studies, no hard-and-fast rules for the minimum sample size given the number of MVs can be generally recommended, especially because MV levels of communality usually cannot be known prior to a study. Instead, larger sample sizes are better, and replication of factor patterns across samples provides greater assurance that an accurate factor structure has been obtained (Fabrigar & Wegener, 2012).

Selection of variables – domain representation and the use of scales, items, or parcels. The selection of MVs to be included in a CFA is a topic that has received more discussion than the selection of observations, but deserves still greater attention (cf. Little, Lindenberger, & Nesselroade, 1999). One key issue is domain representation or coverage. In his initial studies, Thurstone (1938; Thurstone & Thurstone, 1941) attempted to assemble batteries of MVs that spanned a domain of content. For mental abilities, this meant including tests of widely varying content that assessed as many kinds of mental operation as possible. Thurstone had ambitious goals, but hindsight reveals that his initial, large batteries of tests failed to include indicators for many factors that have subsequently been well replicated in multiple studies. Still, as a general rule, careful consideration of the domain of content to be represented in the factor analysis should lead to attempts to ensure that all facets or aspects of the domain are reflected in MVs. When factor analyzing an existing instrument, a researcher might discard items that appear not to reflect the major factors in the instrument or might reformulate the dimensional structure of the instrument (Floyd & Widaman, 1995). In such applications, an investigator may not be able to select the variables analyzed, as this selection is mandated by the aim of the research to evaluate the factor structure of an existing instrument. But experience with an existing scale may lead a researcher to conclude that important aspects of the behavioral domain have, inadvertently or deliberately, been excluded from the existing instrument. In such cases, researchers can supplement an existing instrument with additional pertinent items to see whether this will result in a more adequate, theoretically compelling representation of the domain.

The issue of adequate representation of a domain and all of its facets is a key issue that is often overlooked. When constructing a set of items or other

indicators for a domain of content, the researcher should try to ensure three or more indicators for each aspect or facet of the domain. For example, several facets of Extroversion have been hypothesized, including gregariousness and social potency or assertiveness. Having multiple items for gregariousness and multiple items for assertiveness would enable separate factors for these facets to be identified. However, if gregariousness and assertiveness are closely related facets of Extroversion, then a single Extroversion factor may be sufficient to explain all relations among the gregariousness and assertiveness facets. The issue of whether multiple factors are needed to describe gregariousness and assertiveness or whether a single factor is sufficient is a question that CFA is primed to answer but can do so only if multiple indicators for each facet of a domain are available. If only a single indicator for a particular facet of a domain is included in an assessment battery, a common factor for that domain cannot emerge, because a common latent variable for the domain cannot be separated from the unique factor for that indicator. Thus, a researcher may “miss” or fail to uncover a latent variable for an important facet of a domain because of failure to include more than a single indicator for that facet in analyses.

Another issue worthy of comment is whether the CFA is based on scale (or test) scores, item scores, or parcel scores. The CFA data model assumes that MV scores are linear functions of continuous LV scores. Scale or test scores, which are sums of all items from a scale, appear capable of meeting this assumption. For more than 75 years, however, personality researchers have factor analyzed item scores, which can pose problems for the linear CFA model. Item scores may be binary (0 = fail, 1 = pass) or ordered-categorical in nature (e.g., falling on 1-to-5 “agree-disagree” Likert scale). Artifactual “difficulty” factors can occur when analyzing binary data, a fact that has long been known (Ferguson, 1941; Guilford, 1941; Wherry & Gaylord, 1944). A difficulty factor is an artifactual factor that reflects differences in item difficulty (i.e., item passage rates) rather than item content. Difficulty factors can appear when linear factor analysis models are fit to Pearson correlations among binary items; difficulty factors can be avoided if factor models are fit to matrices of tetrachoric correlations among such items. Problems can also arise in analyses of Likert scale data, and these problems have not been widely recognized in the substantive literature, where standard factor analyses of item data have been conducted with impunity. More recently, many contributions have been made regarding proper ways to analyze binary and ordered-categorical variables in CFA models (e.g., Bock, Gibbons, & Muraki, 1988; McDonald & Ahlawat, 1974; Millsap & Yun-Tein, 2004; Muthén, 1978, 1984; Wirth & Edwards,

2007).

A middle ground on the “scale scores versus items” issue is the use of parcels, where items from a scale are divided into three or four parceled MVs to represent the dimension. A great deal has been written over the past decade and more on the use of parcels, both pro and con. Cattell (1956a, 1956b) was apparently the first to use parcels and coined the term “parcel,” and disputes about parcels remain a topic of current interest. On the pro side, parcel scores are more continuous and are usually distributed more normally than the item scores comprising the parcels, and thus are more likely to meet the linear CFA specification between MVs and LVs. Moreover, the use of parcels avoids the appearance of doublet factors, with only two high loadings. Cattell and Tsujioka (1964) dubbed doublet factors “bloated specifics,” which are attributable to excessive item overlap and do not represent legitimate common factors.

On the con side, many authors have demonstrated that some factors may fail to appear if items are parceled inappropriately. The most important thing to keep in mind when forming parcels is the need to ensure that parcels for a given factor are as representative as possible of the theoretical construct that is the object of measurement. For example, if a scale includes both positively and negatively worded items in order to avoid acquiescence bias, then each parcel derived from the scale should include both positively and negatively worded items. Or, if a scale for a broad construct includes items from several facets of the construct, each parcel might include one or more items from each facet. Research on parcels and their appropriate use in research continues apace. For example, Sterba and MacCallum (2010) illustrated conditions under which results may vary considerably across different ways of assigning scale items to parcels. Little, Cunningham, Shahar, and Widaman (2002) discussed pro and con arguments regarding the use of parcels. More recently, Little, Rhemtulla, Gibson, and Schoemann (2013) and Marsh, Luedtke, Nagengast, Morin, and von Davier (2013) offered updated pro and con arguments on use of parcels.

Implementing CFA

To structure our presentation of how to implement CFA, we discuss a process involving four steps: specification, estimation, evaluation, and readjustment of models.

Specification. Specification of a CFA model involves identifying the number of factors and placement of fixed, free, and constrained parameters. Fixing any model parameter is a constraint that the parameter must take on a specified

value. Two general kinds of constraints can be distinguished: minimally sufficient identification (MSI) constraints and overidentifying restrictions. MSI constraints are used to identify the scale of the LVs (i.e., the mean and variance of each LV) and identify all parameter estimates in the model. At least r^2 MSI constraints must be invoked to identify the covariance structure and r constraints are made to identify the mean structure, where r is the number of LVs. If only $(r^2 + r)$ constraints are made, the resulting model is equivalent to an EFA model; the associated CFA model will have the same chi-square and degrees of freedom as an EFA model with the same number of factors. The CFA model is equivalent to an EFA model but is rotated to an a priori orientation, as will become clear from later discussion.

LVs have no natural scale, and certain MSI identification constraints set or fix the scale of the LVs. One common approach is to fix the variance of each LV to unity, by fixing or constraining $\text{diag}(\Psi) = \mathbf{I}$, and to fix the mean of each LV to zero, or $\alpha = 0$. This approach provides LVs that have properties akin to z-scores, with a mean of zero and a variance of 1.0. A second common identification constraint is to fix one factor loading per LV to unity, which then allows the variance of the LV to be estimated, still with LV means fixed at zero. Let us assume that the former approach – fixing LV variances to unity and LV means to zero – is taken; this accounts for $2r$ of the MSI constraints. The remaining MSI constraints are invoked by fixing certain factor loadings to zero. Consider the following CFA model matrices for six MVs:

$$\begin{aligned} \boldsymbol{\tau} &= \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \\ \tau_6 \end{bmatrix} & \boldsymbol{\alpha} &= \begin{bmatrix} 0^* \\ 0^* \end{bmatrix} & \boldsymbol{\Lambda} &= \begin{bmatrix} \lambda_{11} & 0^* \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ 0^* & \lambda_{42} \\ \lambda_{51} & \lambda_{52} \\ \lambda_{61} & \lambda_{62} \end{bmatrix} \\ \boldsymbol{\Psi} &= \begin{bmatrix} 1^* & \text{symm} \\ \psi_{21} & 1^* \end{bmatrix} & \boldsymbol{\Theta} &= \text{diag} \begin{bmatrix} \theta_{11} \\ \theta_{22} \\ \theta_{33} \\ \theta_{44} \\ \theta_{55} \\ \theta_{66} \end{bmatrix} \end{aligned} \quad (20.7)$$

where the first three MVs (corresponding to the first three rows of Λ) are presumed to be indicators of the first LV and the second three MVs are indicators for the second LV. Because two LVs are shown in these matrices, a total of $(r^2 + r) = 2^2 + 2 = 6$ MSI constraints must be made to identify this model. Two MSI constraints are indicated by the asterisked values of 1 on the diagonal of Ψ , the two remaining MSI constraints for the covariance structure are shown as the asterisked values of 0 in the two columns of Λ , and the two mean structure constraints are shown as asterisked values of 0 for α . The placement of MSI constraints was determined in the following way, which can be generalized to models with any number of factors: (a) select as the “leading” MV indicator for each factor the best indicator of the factor based on prior research, if such can be determined (here, MV1 is the “leading” indicator for LV1, and MV4 is the “leading” indicator for LV2); (b) estimate the factor loading for the “leading” indicator on its factor, but fix at 0 the factor loadings for the “leading” indicator on all other factors; (c) for the “nonleading” indicators of a factor, estimate all loadings for these MVs on all other factors; (d) fix to 1.0 the variances of all LVs in Ψ ; (e) estimate all correlations among LVs; (f) fix to 0 all factor means in α ; and (g) freely estimate diagonal values of Θ , which are variances of unique factors.

The model in Equation 20.7 has $(r^2 + r)$ MSI constraints, and thus is equivalent to an EFA model. The model is simply an EFA model estimated using CFA software and rotating the first LV right through the first MV and the second LV directly through the fourth MV. Because the first factor is collinear with the first MV, the first MV cannot load on any other factors, hence its fixed 0 loadings on other factors. Similar constraints hold for the fourth MV.

Converting the CFA model in Equation 20.7 into a typical, restricted CFA model involves constraining the four underscored factor loadings (i.e., factor loadings λ_{22} , λ_{32} , λ_{51} , and λ_{61}) to be zero but keeping the six other factor loadings (λ_{11} , λ_{21} , λ_{31} , λ_{42} , λ_{52} , and λ_{62}) estimated freely. This involves invoking the second general type of constraint – overidentifying restrictions. Specifically, if the loadings of the second and third MVs on the second LV (i.e., λ_{22} and λ_{32}) and the loadings of the fifth and sixth MVs on the first LV (i.e., λ_{51} and λ_{61}) were fixed at zero, the resulting model would be a CFA model that is a restricted version of the model shown in Equation 20.7. The resulting CFA model would have four more df than the model shown in Equation 20.7 because four fewer parameter estimates are made.

Given the foregoing, one can see that a standard CFA model is simply a restricted version of an EFA model. Typical CFA models are not confirmatory in a strong sense, meaning that a researcher does not test whether particular numerical values for certain model parameters can be confirmed or verified in a new sample. Instead, the typical CFA model involves obtaining best fit estimates of all model parameters, given a particular restricted pattern of fixed and free loadings. Adequate fit of a restricted CFA model thus “confirms,” or supports, the conjecture that a restricted form of factor model is consistent with the data. The restricted form of a CFA model means that it is no longer a freely rotatable solution, as any rotation would fail to retain the pattern of fixed zero loadings in the model. Furthermore, although not often performed, a researcher could test the fit of a CFA model against that of a freely rotatable EFA model with the same number of factors, because the CFA model is nested within the EFA model. This test might be quite informative, providing a statistical test of the worsening of fit associated with the overidentifying restrictions in the CFA model.

Estimation. In CFA, parameters can be estimated using any of a number of methods of estimation. Available methods include maximum likelihood (ML), unweighted least squares, generalized least squares, and a host of variants of these that provide test statistics and standard errors that are robust to violations of normality assumptions (e.g., Browne, 1984; Satorra & Bentler, 1994; Savalei, 2010). Although myriad methods of estimation are available, the more esoteric methods are not frequently used. Instead, standard ML estimation is the method used in the vast majority of applications. Fortunately, ML estimation appears to be relatively robust to at least moderate violations of distributional assumptions, so its use in most studies is probably not a matter of major concern. If violation of assumptions is a concern, we recommend comparing results of ML estimation against those using a robust method of estimation (e.g., Satorra & Bentler, 1994). If important differences in results are obtained, then method of estimation is a matter of concern, and the researcher should justify the choice of method. Technical details of ML estimation are beyond the bounds of the present chapter; interested readers are referred to Bollen (1989) and Jöreskog (1967) for more detail.

In the process of ML estimation, the ML fit function F_{ML} is minimized iteratively to arrive at parameter estimates that describe the data most accurately. If a CFA model fits the data perfectly and therefore perfectly reproduces means and covariances of the MVs, the F_{ML} value will be zero; any form of misfit will

result in a positive value of F_{ML} . If the data satisfy the assumption that manifest variables are distributed in multivariate normal fashion, multiplying F_{ML} by $(N - 1)$ yields a test statistic that is distributed as chi-square. The number of degrees of freedom for this chi-square value is calculated as the difference between the number of sample statistics (MV means and covariances) and the number of parameter estimates in the model. A significant chi-square value provides a statistical basis for rejecting the model as an adequate description of the data.

Evaluation. Evaluation of the fit of a CFA model is typically done on at least two levels: (1) the global fit of the overall CFA model to the data, and (2) a more detailed consideration of each parameter estimate. Global fit of the CFA model can be evaluated using statistical indices or indices of practical fit that are a function of statistical indices. In the preceding paragraphs we discussed the likelihood ratio chi-square test statistic; this test statistic is used as a statistical indicator of model misfit. A significant test statistic provides a statistical basis for rejection of the model tested, indicating significant residual covariation unaccounted for by the model. The test statistic is a direct function of sample size, so trivial model misfit can lead to a significant test statistic if sample size is large, and important model misfit may not lead to a rejectable test statistic if sample size is small. Thus, care must be taken in evaluating the ML test statistic.

Early in the use of CFA, researchers saw the need for measures of model fit that were not closely linked to sample size. One of the first indices, and still a leading one, is the Tucker-Lewis index, or TLI (Tucker & Lewis, 1973), an index also identified as the nonnormed fit index, or NNFI, by some programs. Later, Bentler (1990) proposed the comparative fit index, or CFI (cf. McDonald & Marsh, 1990). The TLI and CFI are both relative fit indices that index the fit of a substantive model in comparison to a null model with no latent variables. The TLI and CFI range generally between 0 and 1.0, with values of 1.0 indicating that a model explains all of the explainable off-diagonal covariation among MVs. As measures of relative goodness of fit, higher TLI and CFI values are better. Research by Hu and Bentler (1999) suggests that the TLI and CFI should be .95 or higher to conclude that a model provides close fit to data.

Another useful measure of global fit is the root mean square error of approximation (RMSEA), first proposed by Steiger and Lind (1980). The RMSEA was formulated to reflect model misfit, so lower values indicate better model fit. The RMSEA is not on any definable scale, but the lowest value the RMSEA can assume is zero, which indicates perfect fit of the model. Any CFA model can fail to fit a correlation or covariance matrix perfectly for at least two

reasons: sampling variability and misfit of the model in the population. The RMSEA is an index of this latter form of misfit. Browne and Cudeck (1993) stated that broad experience with the RMSEA indicated that values of .05 or less indicate close fit of a model to data, .05 to .08 indicate adequate fit, .08 to .10 indicate poor fit, and over .10 indicate unacceptable fit.

One useful adjunct in using the RMSEA is that the sampling variability of the point estimate of the RMSEA is available, so an interval estimate of the RMSEA can be calculated. If the lower limit of the 90% confidence interval (CI) of the RMSEA includes .05, close fit of a CFA model to the data cannot be rejected at $\alpha = .05$. Further, if the lower limit of the 90% CI includes zero, perfect model fit cannot be rejected.

A final useful indicator of global model fit is the standardized root mean squared residual correlation, or SRMR. Hu and Bentler (1999) recommended that the SRMR should be .08 or smaller to conclude that a CFA model fits the data closely.

The second, more detailed level of evaluation of model fit is conducted by considering each parameter estimate in a CFA model relative to its standard error (SE). A critical ratio (CR), distributed as an asymptotic z-ratio, can be formed by dividing a parameter estimate by its SE. Common practice is to deem any parameter with a CR of 2.0 or larger to differ significantly from zero at the .05 level, and the *p*-level for each CR is reported by most structural modeling programs. Thus, on a statistical basis, the evaluation of model parameters is highly structured.

Parameter estimates in a CFA model should also be evaluated with regard to their magnitude, supplementing a statistical evaluation with one based on practical importance. No hard-and-fast rules are available here, but simulation studies have often used .4, .6, and .8 standardized factor loadings to represent low, medium, and high levels of communality, respectively. With regard to correlations among LVs, Cohen (1988) used .10, .30, and .50 to indicate small, medium, and large correlations, although these recommended levels were for correlations among MVs, so may need to be adjusted for correlations among LVs. The preceding values are offered as approximate levels of small, medium, and large values when interpreting the magnitude of parameter estimates in CFA models.

Readjustment. Readjustment of a CFA model consists of the respecification of fixed and/or free parameter estimates. Respecification can be done on either

of two grounds: (1) a priori and (2) empirical. A priori respecifications might be pursued to compare competing conjectures under different models for a given set of data. For example, in prior research, certain investigators may have found evidence that a three-factor structure underlies items on a given scale, whereas the developer of the scale sought to assess just two factors. Alternative CFA models could be formulated that were consistent with each of these competing schemes, and the relative fit of these alternate a priori models could be compared.

Researchers can also use empirical bases for model respecification, typically to improve model fit to the data. Most structural modeling programs report modification indices (MIs), which estimate the change in chi-square if a particular parameter were estimated, rather than fixed in the current model. Experts on CFA differ with regard to their recommendations concerning MIs. Some experts advise never to base model respecification on MIs because of the capitalization on chance that can readily occur. Bolstering this position is Monte Carlo simulation work on specification searches, which has generally concluded that such searches rarely arrive at the model that generated the data (e.g., MacCallum, 1986). Other experts (e.g., Sörbom, 1989) see clear value in using MIs to aid model respecification, as these indices reflect ways the data are informing the researcher of a need to account for particular patterns in data. If MIs are used to modify the specification of a model, most experts urge that these modifications be cross-validated in an independent sample to verify their replicability. Moreover, if sample size is large enough, a researcher could divide the sample randomly into two groups, using data from one group to develop a best fitting model, which could be cross-validated in the second group.

Two Empirical Examples

Personality data. To illustrate the use of CFA, empirical examples are helpful. In Table 20.1, correlations among 15 parcels of items from the Big Five Inventory (BFI) (John, Donahue, & Kentle, 1991) are shown for two samples, one sample of females ($N = 600$) and one sample of males ($N = 600$), along with means and SDs for each of the parcels for each sample. The BFI is a 44-item instrument to measure the Big 5 personality constructs and contains 8 to 10 items per dimension. The Extraversion scale has 8 items, so the three parcels for Extraversion represent two 3-item parcels and one 2-item parcel. Both the Agreeableness and Conscientiousness scales have 9 items, so the parcels for these dimensions are three 3-item parcels. Details on parcel construction are

available at <http://psychology.ucdavis.edu/labs/Grimm/personal/downloads.html>.

Table 20.1. Correlations Among 15 Item Parcels from the Big Five Inventory, with Means and SDs, in Two Samples – Females and Males

Variable	Variable														
	Ext1	Ext2	Ext3	Agr1	Agr2	Agr3	Con1	Con2	Con3	Neu1	Neu2	Neu3	Ope1	Ope2	Ope3
Ext1		.724	.723	.058	.082	.107	.126	.193	.149	-.272	-.229	-.070	.219	.196	.121
Ext2	.733		.729	.037	.031	.079	.108	.197	.120	-.228	-.188	-.017	.223	.222	.132
Ext3	.752	.727		.093	.087	.127	.089	.157	.113	-.162	-.141	.008	.195	.255	.168
Agr1	.068	.023	.093		.680	.625	.086	.076	.167	-.202	-.248	-.271	.200	.114	.093
Agr2	.159	.134	.149	.605		.634	.174	.165	.245	-.139	-.162	-.193	.160	.100	.068
Agr3	.152	.123	.203	.513	.557		.141	.187	.178	-.205	-.249	-.257	.093	.053	.020
Con1	.046	-.004	-.063	.123	.121	.144		.607	.538	-.226	-.151	-.100	.074	-.015	.047
Con2	.082	.095	.046	.078	.174	.203	.540		.564	-.264	-.228	-.122	.067	.076	.042
Con3	-.042	-.037	-.079	.118	.105	.108	.577	.569		-.232	-.186	-.137	.147	.101	.127
Neu1	-.223	-.180	-.145	-.252	-.250	-.238	-.114	-.209	-.184		.725	.647	-.119	-.089	-.045
Neu2	-.229	-.165	-.147	-.236	-.309	-.297	-.091	-.202	-.161	.681		.680	-.043	-.084	-.041
Neu3	-.048	-.009	-.007	-.283	-.343	-.284	-.073	-.130	-.128	.606	.624		-.083	-.035	-.064
Ope1	.151	.138	.157	.142	.069	-.011	.033	.020	.006	.010	-.002	.033		.633	.677
Ope2	.126	.096	.132	.074	.012	-.020	-.044	.013	-.027	.051	.013	.088	.629		.554
Ope3	.013	.027	.024	.058	.065	-.107	-.001	.000	.012	.070	.046	.038	.594	.522	
Female <i>M</i>	3.268	3.353	3.343	3.640	3.848	3.768	3.918	3.242	3.770	2.863	3.256	3.232	4.254	4.073	3.973
Female <i>SD</i>	0.925	0.976	1.015	0.804	0.772	0.792	0.840	0.865	0.730	0.879	0.896	0.951	0.667	0.792	0.836
Male <i>M</i>	3.071	3.127	3.022	3.495	3.772	3.609	3.765	3.192	3.686	2.448	2.830	2.842	4.266	4.031	4.088
Male <i>SD</i>	0.899	0.964	0.976	0.891	0.838	0.865	0.823	0.858	0.737	0.945	0.946	1.022	0.683	0.789	0.791

Note: Tabled values are correlations for females ($N = 600$) below diagonal and from males ($N = 600$) above the diagonal, with means and SDs as noted. Parcel designations: Ext = Extraversion, Agr = Agreeableness, Con = Conscientiousness, Neu = Neuroticism, Ope = Openness.

For this first example, we selected the 15 MVs corresponding to the five personality factors for the female sample. Thus, we used the correlations below the diagonal of Table 20.1 along with the means and SDs for the female sample in these analyses. We fit a five-factor solution, assuming that the three parcels for each LV would load only on the hypothesized factor. The a priori model had a significant likelihood ratio test statistic, $\chi^2(24, N = 600) = 223.96, p < .0001$, suggesting rejection of the model, but sample size was large leading to a very powerful test statistic. Moreover, all practical fit indices were adequate. Specifically, both the TLI = .950 and CFI = .962 were at or above the .95 threshold; the RMSEA of .055 was above the .05 cut-off, but its CI [.046, .063] included .05, so close fit could not be rejected ($p = .17$); and the SRMR of .041 was well below the recommended cut-off of .08. Hence, we consider the a priori

five-factor model to have fairly close fit to the data.

However, one large MI was found, for the third Neuroticism parcel (Neu3) on the Extroversion factor. When we freely estimated this loading, all indices of model fit improved. The likelihood ratio statistic was still significant, $\chi^2 (24, N = 600) = 195.97, p < .0001$, but represented a substantial improvement in fit over the a priori model for the one additional parameter estimate, $\Delta\chi^2 (1, N = 600) = 27.99, p < .0001$. Further, the TLI = .959 and CFI = .969 were improved; the RMSEA of .050 now fell right at the criterion for close fit, so its CI [.041, .058] fell equally on either side, and the test of close fit was nonsignificant ($p = .51$); and the SRMR of .037 was reduced, all signifying closer fit of this model to the data.

Parameter estimates for the model are shown in [Table 20.2](#). The factor loadings reported are raw-score, or covariance metric, loadings; all were significant with critical ratios over 6. The corresponding standardized loadings (not shown) for hypothesized loadings were large, ranging from .70 to .87. The single deviation from a priori specification was the fairly minor loading of Neu3 on Extraversion, which was significant but small in magnitude. As shown, Agreeableness correlated significantly, but at low levels with Extraversion and Conscientiousness, Neuroticism was negatively correlated with the first three factors, and remaining among factors tended to be quite small in magnitude and nonsignificant.

Table 20.2. Results of Confirmatory Factor Analysis for 15 MVs from the BFI, Female Sample (N=600)

MV	τ	Factor					Θ	h^2
		Ext	Agr	Con	Neu	Ope		
Ext1	3.27 (.04)	.81 (.03)	.0* (—)	.0* (—)	.0* (—)	.0* (—)	.21 (.02)	.77
Ext2	3.35 (.04)	.82 (.03)	.0* (—)	.0* (—)	.0* (—)	.0* (—)	.28 (.02)	.71
Ext3	3.34 (.04)	.87 (.04)	.0* (—)	.0* (—)	.0* (—)	.0* (—)	.27 (.03)	.74
Agr1	3.64 (.03)	.0* (—)	.59 (.03)	.0* (—)	.0* (—)	.0* (—)	.30 (.03)	.54
Agr2	3.85 (.03)	.0* (—)	.63 (.03)	.0* (—)	.0* (—)	.0* (—)	.20 (.02)	.66
Agr3	3.77 (.03)	.0* (—)	.55 (.03)	.0* (—)	.0* (—)	.0* (—)	.33 (.03)	.48
Con1	3.91 (.03)	.0* (—)	.0* (—)	.62 (.03)	.0* (—)	.0* (—)	.33 (.03)	.54
Con2	3.24 (.04)	.0* (—)	.0* (—)	.64 (.04)	.0* (—)	.0* (—)	.34 (.03)	.54
Con3	3.77 (.03)	.0* (—)	.0* (—)	.57 (.03)	.0* (—)	.0* (—)	.21 (.02)	.61
Neu1	2.86 (.04)	.0* (—)	.0* (—)	.0* (—)	.71 (.03)	.0* (—)	.27 (.02)	.65
Neu2	3.26 (.04)	.0* (—)	.0* (—)	.0* (—)	.75 (.03)	.0* (—)	.25 (.02)	.69
Neu3	3.23 (.04)	.17 (.03)	.0* (—)	.0* (—)	.76 (.04)	.0* (—)	.36 (.03)	.60
Ope1	4.25 (.03)	.0* (—)	.0* (—)	.0* (—)	.0* (—)	.57 (.03)	.12 (.02)	.73
Ope2	4.07 (.03)	.0* (—)	.0* (—)	.0* (—)	.0* (—)	.59 (.03)	.28 (.02)	.55
Ope3	3.97 (.03)	.0* (—)	.0* (—)	.0* (—)	.0* (—)	.58 (.03)	.36 (.03)	.48
Factor Correlations								
Ext		1.0* (—)						
Agr		.19 (.05)	1.0* (—)					
Con		.00 (.05)	.23 (.05)	1.0* (—)				
Neu		-.26 (.05)	-.46 (.04)	-.24 (.05)	1.0* (—)			
Ope		.17 (.05)	.08 (.05)	.01 (.05)	.04 (.05)	1.0* (—)		

Note: Tabled values are raw-score intercepts (τ), factor loadings, unique factor variances (Θ), and factor correlations, with standard errors (*SEs*) in parentheses. All parameter estimates had z-ratios greater than 6.0, except factor correlations, several of which were nonsignificant. Parcel and factor designations: Ext = Extroversion, Agr = Agreeableness, Con = Conscientiousness, Neu = Neuroticism, and Ope = Openness.

* Asterisk values fixed at reported values to identify model, so *SEs* are not available.

A multitrait-multimethod example. In a seminal article, Campbell and Fiske (1959) discussed convergent and discriminant validation using the multitrait-multimethod (MTMM) matrix. Convergent validation was shown if measures purportedly of the same construct but assessed with different methods had high correlations with one another, and discriminant validation was supported if measures of different constructs had lower levels of correlation. With the advent of CFA, researchers quickly saw the utility of restricted factor models for evaluating trends in MTMM matrices. That is, a researcher could specify combinations of trait and method factors to test the Campbell and Fiske notions statistically. Note that correlations are allowed among trait factors and among method factors, but correlations between trait and method factors are fixed at

zero to ensure identification of model parameters. This model is often termed the *correlated trait-correlated method* (CT-CM) model. Jöreskog (1971b) was one of the first to fit CFA models to MTMM data, analyzing one of the Campbell and Fiske matrices.

To reconcile discrepancies in work to that date, Widaman (1985) offered a taxonomy of CFA models for MTMM data. Widaman identified four possible trait factor structures – (1) an absence of trait factors, (2) trait factors that were perfectly correlated, (2') trait factors that were perfectly uncorrelated, and (3) trait factors that were freely correlated – and four possible method factor structures – (A) an absence of method factors, (B) method factors that were perfectly correlated, (B') method factors that were perfectly uncorrelated, and (C) method factors that were freely correlated. Cross-classifying the four trait and four method structures resulted in a taxonomy of 16 potential CFA models for MTMM data. The CT-CM model fit first by Jöreskog (1971b) is one of the 16 models in the Widaman's taxonomy, a model identified as Model 3C, given the combination of trait and method structures. A short time later, Marsh (1989), building on work by Kenny (1976), offered a fifth possible method structure – allowing covariances among unique factors for all indicators for a given method. This led to the *correlated trait-correlated uniqueness*, or CT-CU, model. At the turn of the century, Eid (2000) argued that deleting one method factor would correct common convergence problems in the CT-CM model. Deleting one method factor thus led to the *correlated trait-correlated method-minus-1*, or CT-C(M-1), model. A recent summary of multimethod approaches to research is contained in Eid and Diener (2006), and Eid, Lischetzke, and Nussbeck (2006) provide an up-to-date summary of different structural modeling approaches to evaluating MTMM data.

An empirical MTMM matrix is shown Table 20.3, which shows relations among nine MVs assessing child personality traits for a sample of 526 families. In this table, the traits assessed are the higher-order constructs of Positive Emotionality, Negative Emotionality, and Constraint in the Tellegen (1982) model, and raters – father, mother, and child – are the methods. The convergent validities, shown in boldface, are high between father and mother ratings, ranging between .51 and .68, and the child self-ratings exhibited lower but still substantial levels of convergent validity, ranging between .26 and .44.

Table 20.3. Multitrait-Multimethod Matrix for Measures of Three Child Personality Traits Rated by Father, Mother, and Child

Variable	Variable								
	Father			Mother			Child		
	Pos	Neg	Con	Pos	Neg	Con	Pos	Neg	Con
Father Pos									
Father Neg	<i>-.488</i>								
Father Con	<i>.445</i>	<i>-.536</i>							
Mother Pos	.619	<i>-.323</i>	<i>.295</i>						
Mother Neg	<i>-.336</i>	.514	<i>-.399</i>	<i>-.473</i>					
Mother Con	<i>.270</i>	<i>-.356</i>	.683	<i>.342</i>	<i>-.537</i>				
Child Pos	.340	<i>-.135</i>	<i>.050</i>	.377	<i>-.151</i>	<i>-.021</i>			
Child Neg	<i>-.206</i>	.260	<i>-.262</i>	<i>-.258</i>	.326	<i>-.316</i>	<i>-.293</i>		
Child Con	<i>.071</i>	<i>-.175</i>	.394	<i>.054</i>	<i>-.183</i>	.443	<i>-.071</i>	<i>-.240</i>	
<i>M</i>	10.590	7.532	9.964	10.534	7.191	10.112	2.015	1.294	1.829
<i>SD</i>	1.608	1.706	1.685	1.804	1.895	1.884	0.458	0.595	0.488

Note: Tabled values are estimated correlations based on FIML estimation for a sample of families ($N = 526$), with means and SDs as noted. Pos = Positive Emotionality, Neg = Negative Emotionality, and Con = Constraint. Convergent validities are shown in boldfaced type, correlations among measures using the same method in italic type.

The first CFA model fit to the data was a model with correlated traits only. As shown in Figure 20.1A, this model posited the existence of three trait factors: the Positive Emotionality factor had direct effects on the father, mother, and child ratings of child positive emotionality; the Negative Emotionality factor had direct effects on father, mother, and child ratings of child negative emotionality; and the Constraint factor had direct effects on father, mother, and child ratings of child constraint. This model had poor fit to the data, with a significant test statistic, $\chi^2(24, N = 526) = 248.04, p < .0001$, and poor practical fit indices, with TLI and CFI below .85 and an RMSEA over .10, indicating clear rejection of close fit, $p < .0001$. Parameter estimates for this model are shown in Figure 20.1A, where the father and mother ratings had very large loadings on the three trait factors, the child ratings had rather low loadings, and the three trait factors were correlated at high levels, indicating relatively poor discriminant validity.

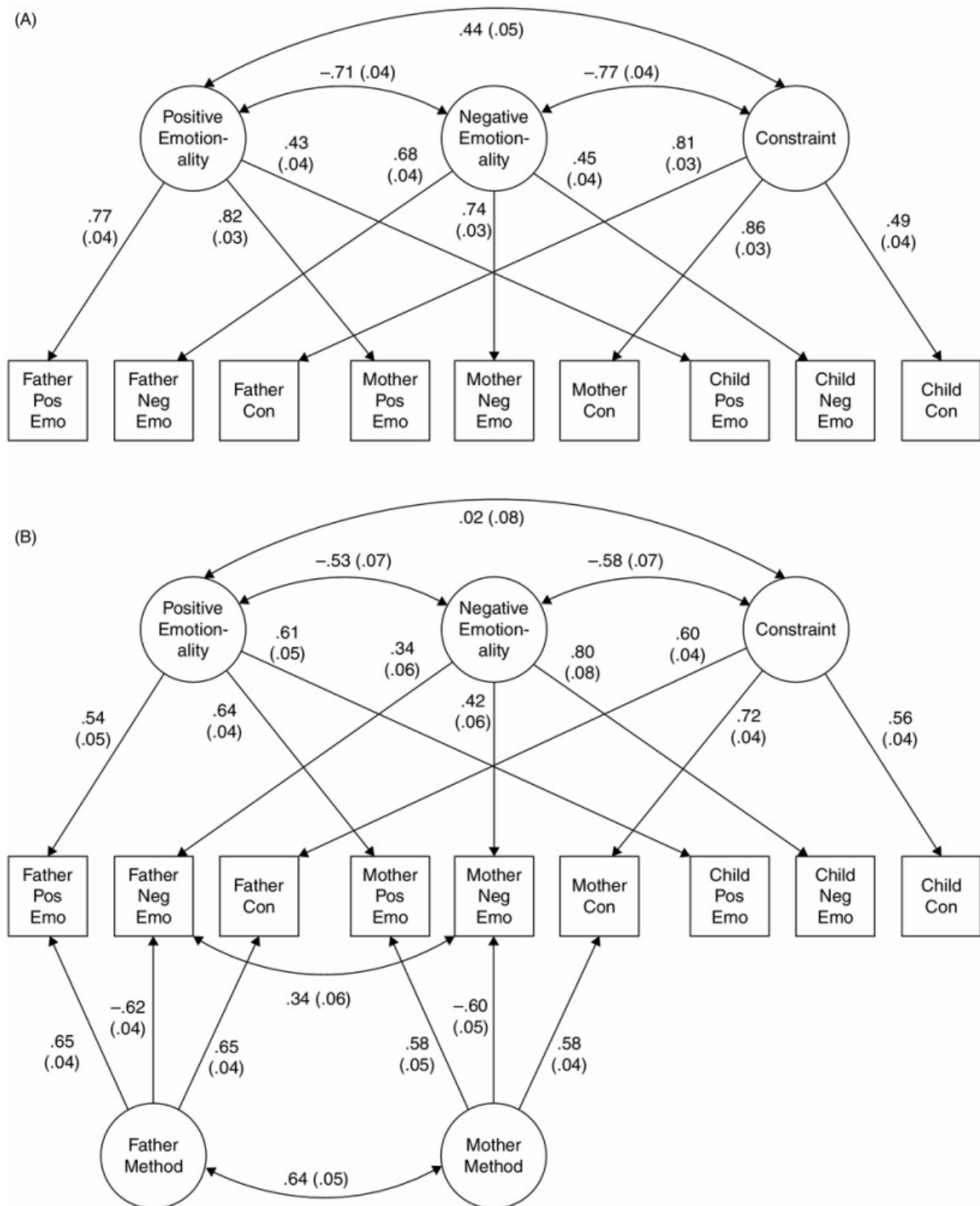


Figure 20.1. Results of multitrait-multireporter analysis of ratings of child trait levels: (A) Model with three trait factors only. (B) Model with three trait factors and two method factors. (Values are standardized parameter estimates, with SEs

in parentheses. Pos emo = Positive Emotionality; Neg emo = Negative Emotionality; Con = Constraint).

A second model fit to the data, which added three correlated method factors, did not converge, due to the low correlations among the child rating MVs, as shown in [Table 20.3](#). Thus, a modified model – with method factors for the father ratings and mother ratings, but not for child ratings – fit the data much better. Here, the Father method factor had direct effects on the three ratings made by fathers and the Mother method factor had direct effects on the three ratings made by mothers. One further modification was made, allowing a covariance between unique factors for father and mother ratings of negative emotionality. The final model had very close fit to the data, with χ^2 (16, $N = 526$) = 25.55, $p = .06$, and very good practical fit index values, with TLI = .984 and CFI = .993 and an RMSEA = .034 (95% CI [.000, .057]), $p(\text{close fit}) = .86$. Parameter estimates for this model are shown in [Figure 20.1B](#) and offer a marked contrast with those in [Figure 20.1A](#). Specifically, in the model in [Figure 20.1B](#), the child ratings had factor loadings on trait factors that were either approximately equal to the father and mother loadings (for Positive Emotionality and Constraint) or much higher than the father and mother loadings (for Negative Emotionality). This latter result is particularly important, as Negative Emotionality is the most private of the three traits, and the child should, based on theory, be a more informed rater of this trait. Finally, the father and mother method factors had substantial loadings, indicating strong halo effects in their ratings. Squaring factor loadings to obtain estimates of variance explained, from 33% to 42% of the variance on each of the father and mother rating variables was attributable to method, or halo, effects that were unrelated to the trait constructs under consideration.

This MTMM example illustrates vividly the benefits accruing to the use of CFA modeling in several ways. First, results for the final model show that a substantial amount of variance in parent ratings of their child's personality can be contaminated with construct-irrelevant bias – positive and negative – regarding their offspring. Second, comparing the models in [Figures 20.1A](#) and [20.1B](#), conclusions about the relations among trait constructs may be substantially in error if the contaminating effects of bias are not removed. Finally, the analysis highlights the flexibility of the CFA approach, which is far more adaptable than EFA for answering questions of key theoretical and empirical import.

Factorial Invariance in CFA Models

Theoretical requirements. Having shown the utility of CFA models to capture dimensional structures of theoretical interest, we turn now to evaluating measurement invariance with such models. Groups can be identified by subsetting data as a function of person characteristics such as sex or ethnic status. A crucial scientific question quickly arises: Are factors identified in one group the same factors as those identified in other groups? The answer to this question is crucial; if factors differ in their nature across groups, then cross-group comparisons on the factors have no meaning or interpretation. This question of measurement invariance at the level of factors, or LVs, has been studied under the rubric of factorial invariance.

Work on factorial invariance can be traced back over the past 60 years or more, with key contributions by Meredith (1964a, 1964b, 1993), Jöreskog (1971a), and Horn, McArdle, and Mason (1983), among others. But factorial invariance and its implications continue to be matters of great interest. In addition to the preceding sources, interested readers should consult further literature, including Browne and Arminger (1995), Cheung and Rensvold (1999), Levy and Hancock (2007), Little (1997), McArdle (1996), McArdle and Cattell (1994), and Millsap and Meredith (2007).

Concerns about measurement invariance often arise when investigating the differential validity of psychological tests with persons from different ethnic groups. Cleary (1968) offered a series of statistical tests using multiple regression analysis to determine whether intercept bias and/or slope bias existed in the use of test scores when evaluating students from different ethnic groups for college admission. Suppose that the intercept and slope for a test used as a predictor of an outcome such as college success (e.g., GPA) are identical for a reference group (e.g., European-American students) and a comparison group (e.g., African-American students). In this scenario, the test is viewed as being unbiased for use as a selection device, because a given test score translates into the same predicted criterion score in each group. If fewer applicants from the comparison population were selected using such a test, one could justify the result by claiming that the lower level of selection was attributable to lower levels of scores obtained by members of the comparison population on an unbiased test.

However, if group differences are found on the intercept, the slope, or both intercept and slope, then the test is biased, and a given test score translates into

different predicted criterion scores in the two groups. If slopes for a predictor differ significantly across groups, then group is a moderator of the relation between the predictor and the criterion, and slope bias is present. Further, if the group main effect is significant when the predictor variable takes on a value of zero, intercept bias is present. If intercept bias, slope bias, or both intercept and slope bias occur, use of the test as a “gold standard” predictor is suspect, given the different predictor functions across groups. Recently, Millsap (2011) showed that, in addition to inequality of intercept and slope, inequality of residual variance on the criterion can also affect selection decisions, so equality of residual variance on the criterion should also be investigated.

Pursuing measurement invariance with CFA, a model akin to Equations 20.3 and 20.4 can be written as:

$$E(Y_g) = \tau_g + \Lambda_g \alpha_g \quad (20.8)$$

$$E(Y_g Y_g') = \Sigma_{YY_g} = \Lambda_g \Psi_g \Lambda_g' + \Theta_g \quad (20.9)$$

where the g subscript ($g = 1, \dots, G$) is an indicator of group membership, α_g is a vector of mean differences for group g , and other symbols were defined above. The g subscript on the matrices in Equations 20.8 and 20.9 indicates that different parameter estimates may be present for corresponding parameters across groups. However, the g subscript on a given matrix can be deleted if parameter estimates are constrained to numerical invariance across groups.

Widaman and Reise (1997) synthesized prior work by Meredith (1993) and Horn *et al.* (1983) and identified four levels of factorial invariance, which were identified as:

1. *configural invariance*, or invariance of the pattern of fixed and free factor loadings in Λ ;
2. *weak factorial invariance*, or invariance of estimated loadings in Λ ; because factor loadings are raw score regression weights for predicting manifest variables from latent variables, weak factorial invariance ensures that a one-unit change on a latent variable will translate into the identical predicted change in the particular manifest variable in all groups;
3. *strong factorial invariance*, or invariance of estimated loadings in Λ and measurement intercepts in τ ; because measurement intercepts in

τ are intercepts in linear models predicting manifest variables, strong factorial invariance means that a given score on a latent variable will translate into the identical predicted score on a particular manifest variable in all groups; and

4. *strict factorial invariance*, or invariance of estimated loadings in Λ , measurement intercepts in τ , and unique factor variances in Θ ; because unique factor variances are residual variances when predicting manifest variables from latent variables, strict factorial invariance ensures that the residual variance when predicting a particular manifest variable from the latent variables is the same across groups, indicating homogeneity of residual variance across groups.

The preceding factorial invariance models (1) through (4) represent increasingly constrained models. That is, the configural invariance model is the least constrained of the models, and the weak factorial invariance model imposes across-group invariance constraints on factor loadings in the configural invariance model. Because of this, the weak factorial invariance model is nested within the configural invariance model, and the difference in chi-square values for the two models is distributed as a chi-squared variate with degrees of freedom equal to the difference in degrees of freedom for the two models. Thus, one can test whether the change in fit of the model to the data reflects a significant worsening of statistical fit associated with invariance constraints on factor loadings, and one can also evaluate changes in practical fit indices. Similar comparisons can be made with regard to the strong and strict factorial invariance models. That is, each model on the preceding list is nested within the model preceding it, and imposing a given set of invariance constraints can be evaluated with regard to change in statistical fit and change in practical fit.

The key notion of measurement or factorial invariance models is this: If intercept and slope coefficients in a linear model relating predictor variable(s) to a criterion variable are invariant (or identical) across groups, then measurement or factorial invariance holds. If coefficients are invariant across groups, then the predictor variables – the latent variables in CFA models – exhibit no bias in prediction. Thus, work on factorial invariance represents a translation of the concepts regarding test bias in Cleary's (1968) regression models into the multiple-group CFA arena. Note that, in CFA models, the manifest Y variables are the outcome or criterion variables (cf. Equation 20.1), and the LVs are the predictor variables. As a result, the slope coefficients relating predictors to outcomes in CFA models are embodied in the factor loadings in Λ (i.e., factor

loadings are regression weights for predicting MVs from LVs), and the intercepts are the values in the τ vector. Thus, if strong factorial invariance is satisfied and estimates in the Λ and τ matrices are invariant across groups, then the LVs are unbiased predictors of MVs, and mean differences across groups on MVs are accounted for by group mean differences on the LVs. Further, if strict factorial invariance holds, all between-group differences in means and variances on MVs are a function of between-group differences in means and variances on LVs. In a mathematical sense, the latent variables explain or represent all group differences on MVs if strict factorial invariance holds for a set of data.

Under the weak factorial invariance model, slopes relating LVs to MVs are invariant across groups, and between-group differences in variances on LVs and in covariances among LVs are identified in an invariant fashion. Thus, across-group differences in LV variance-covariance matrices can be studied to see if, for example, males exhibited more or less variance on each factor than females do. But if strong or strict factorial invariance holds across groups, the full regression function relating LVs to MVs – both intercepts and slopes – are invariant across groups, and a researcher has putative evidence that the same factors or LVs have been found across groups. Given this, a researcher may investigate group differences in means on the LVs, in addition to differences in variances on the LVs and in covariances among the LVs. Group differences in means on LVs or on variances and/or covariances among LVs can be evaluated in ways analogous to those for testing differences among factorial invariance models, discussed earlier. That is, one can impose across-group invariance constraints on particular sets of parameter estimates and evaluate whether statistical fit worsens significantly and whether practical fit is affected to any great degree. But, to be clear, if only configural invariance across groups holds for a set of data, group differences in mean levels or in variances and covariances of LVs have no clear interpretation. This underscores the need to establish strong or strict factorial invariance for group differences in LV means and LV variance-covariance matrices to have meaningful interpretation.

Empirical example. The preceding approach to testing factorial invariance can be applied to the 15 variables reflecting Big 5 personality factors from the samples of females and males contained in [Table 20.1](#). First, we fit a nonoverlapping simple structure solution in both groups, with the three key indicators of each factor having the only nonzero loadings on each factor. This model, Model 1, is the configural invariance model because the same pattern of fixed and free factor loadings was specified in each group. Model 1 had

generally adequate fit. As shown in Table 20.4, the likelihood ratio test was significant, $\chi^2(160) = 454.62$, $p < .001$, suggesting rejection of the model. However, total sample size ($N = 1200$) was rather large, so the likelihood ratio test had very high power to detect model misfit. Consistent with this interpretation, practical fit indices were quite satisfactory, with an RMSEA = .050, both CFI and TLI over .95, and SRMR below .05. In both samples, a modification index indicated that the third Neuroticism parcel should load on the Extraversion factor; finding this in both samples is a form of cross-sample replication or cross-validation of the need for this loading. Therefore, we fit a second configural invariance model that additionally estimated this loading in each sample. The resulting model, Model 1a, had better statistical fit to the data, $\chi^2(158) = 387.98$, $p < .001$, and both CFI and TLI were now over .96. Because Model 1 was nested within Model 1a, we could compare the statistical fit of the models. Model 1a showed a strong improvement in the statistical index of fit over Model 1, $\Delta \chi^2(2) = 66.64$, $p < .001$, supporting the addition of the extra loading in each sample. Thus, in Model 1a, each of the five personality factors had loadings for the three parcels constructed for the factor, and the Extroversion factor had a single non-hypothesized loading by the third Neuroticism parcel. This latter deviation from perfect nonoverlapping simple structure is a fairly minor matter, and not surprising in a highly restricted model such as Model 1a.

Table 20.4. Indices of Fit for Alternative Factorial Invariance Models Evaluating Invariance of the Big 5 Personality Traits Across Females and Males

Model	χ^2	df	RMSEA (CI)	CFI	TLI	SRMR
1. Configural Invariance	454.62	160	.050 (.049, .061)	.963	.952	.041
1a. Model 1 plus one extra loading	387.98	158	.049 (.043, .055)	.971	.962	.036
2. Weak factorial invariance	396.78	169	.047 (.041, .053)	.971	.965	.038
3. Strong factorial invariance	434.08	179	.049 (.043, .055)	.968	.963	.039
4. Strict factorial invariance	451.68	194	.047 (.041, .053)	.968	.965	.043
5. Model 4 plus invariance of LV covariances	497.68	209	.048 (.043, .053)	.964	.964	.065
6. Model 4 plus invariance of LV means	605.76	199	.058 (.053, .064)	.949	.946	.060

Note: See text for distinctions among alternative models. Female $N = 600$, male $N = 600$. Column headings: χ^2 = likelihood ratio chi-square; df = degrees of freedom; RMSEA (CI) = root mean square error of approximation, with its 90% confidence interval in parentheses; CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean square residual.

Model 2 was the weak factorial invariance model, which was identical to Model 1a but invoked invariance across groups of factor loadings in Λ . The fit of Model 2 was good, with a significant χ^2 value, but unchanged or even improved practical fit indices. For example, both the RMSEA and the TLI improved moving from Model 1a to Model 2. Further, Model 2 showed a nonsignificant worsening of fit associated with the invariance constraint on factor loadings, $\Delta\chi^2(11) = 8.80, ns$. The good fit of Model 2 demonstrates that a model in which all regression weights for predicting MVs from LVs are invariant across groups is a satisfactory representation of the data, supporting slope invariance of the prediction of MVs from the LVs in the model.

The next model, Model 3, was the strong factorial invariance model, which was identical to Model 2 except that measurement intercepts were now also constrained to invariance across groups. The fit of Model 3 was satisfactory, with a significant χ^2 test. Compared with Model 2, Model 3 showed a significant worsening of fit associated with the invariance constraint on intercepts, $\Delta\chi^2(10) = 37.30, p < .0001$. But, owing to interpretive benefits of the strong factorial invariance model relative to the weak factorial invariance model and the fact that practical fit indices worsened only slightly, we accepted Model 3 as an adequate representation of the data. Indeed, the point estimate of the RMSEA (.049) fell below the .05 cut-off value signifying close fit of a model to data, so Model 3 is a well-fitting model. In Model 3, the full regression functions for predicting MVs from LVs – both the intercept and slope(s) for predicting each MV from the LVs in the model – are the same across groups, so a particular score on a given LV would translate into the same predicted MV score whether the participant was male or female.

Our next model, Model 4, was the strict factorial invariance model, which was identical to Model 3 except that invariance of unique variances was also enforced. The fit of Model 4 was quite good. The χ^2 test was still significant, but the CFI was unchanged and the RMSEA and TLI both improved. Supporting the invariance constraints on unique variances, the fit of Model 4 was not significantly worse than that of Model 3, $\Delta\chi^2(15) = 17.60, ns$. Thus, the strict factorial invariance model was accepted as the optimal model for these data. In Model 4, in addition to intercept and slope invariance, the male and female samples exhibited unique factor variances – or variances of MV residuals – that did not differ across groups, suggesting that error variability about the regression lines was comparable across groups.

Parameter estimates from Model 4 are shown in Table 20.5. The three primary indicators of each factor had large raw-score loadings, ranging from .55 to .87. The corresponding standardized factor loadings (not shown) were quite large, ranging from .72 to .87; standardized loadings departed from invariance across groups because groups differed in LV variances in Model 4, but all standardized loadings in each group were large. The three a priori loadings on each factor were significant, with z-ratios over 30. The single added loading – of the Neu3 parcel on the Extraversion factor – was relatively small (.18) but significant ($z > 7$, $p < .0001$). With regard to measurement model parameters in the $\boldsymbol{\tau}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\Theta}$ matrices, all estimates were significant and accurately estimated, with small SEs.

Table 20.5. Results for the Strict Factorial Invariance Model for Evaluating Invariance of the Big 5 Personality Traits Across Females and Males

MV	$\boldsymbol{\tau}$	Factor					$\boldsymbol{\Theta}$
		Ext	Agr	Con	Neu	Ope	
Ext1	3.29 (.04)	.80 (.03)	.0* (–)	.0* (–)	.0* (–)	.0* (–)	.21 (.01)
Ext2	3.36 (.04)	.84 (.03)	.0* (–)	.0* (–)	.0* (–)	.0* (–)	.27 (.02)
Ext3	3.31 (.04)	.87 (.03)	.0* (–)	.0* (–)	.0* (–)	.0* (–)	.27 (.02)
Agr1	3.63 (.03)	.0* (–)	.0* (–)	.0* (–)	.0* (–)	.0* (–)	.28 (.02)
Agr2	3.87 (.03)	.0* (–)	.61 (.03)	.0* (–)	.0* (–)	.0* (–)	.22 (.02)
Agr3	3.75 (.03)	.0* (–)	.60 (.03)	.0* (–)	.0* (–)	.0* (–)	.32 (.02)
Con1	3.89 (.03)	.0* (–)	.56 (.03)	.62 (.03)	.0* (–)	.0* (–)	.31 (.02)
Con2	3.27 (.03)	.0* (–)	.0* (–)	.66 (.03)	.0* (–)	.0* (–)	.31 (.02)
Con3	3.77 (.03)	.0* (–)	.0* (–)	.55 (.03)	.0* (–)	.0* (–)	.24 (.02)
Neu1	2.85 (.03)	.0* (–)	.0* (–)	.0* (–)	.72 (.03)	.0* (–)	.26 (.02)
Neu2	3.25 (.04)	.0* (–)	.0* (–)	.0* (–)	.75 (.03)	.0* (–)	.23 (.02)
Neu3	3.27 (.04)	.18 (.02)	.0* (–)	.0* (–)	.75 (.03)	.0* (–)	.38 (.02)
Ope1	4.25 (.03)	.0* (–)	.0* (–)	.0* (–)	.0* (–)	.58 (.02)	.11 (.01)
Ope2	4.04 (.03)	.0* (–)	.0* (–)	.0* (–)	.0* (–)	.57 (.03)	.30 (.02)
Ope3	4.02 (.03)	.0* (–)	.0* (–)	.0* (–)	.0* (–)	.59 (.03)	.32 (.02)
Females		Factor Covariances					
Ext	.0* (–)	1.0* (–)	.18	.01	–.26	.16	
Agr	.0* (–)	.18 (.05)	1.0* (–)	.23	–.46	.08	
Con	.0* (–)	.01 (.05)	.23 (.05)	1.0* (–)	–.24	.01	
Neu	.0* (–)	–.26 (.05)	–.46 (.04)	–.24 (.05)	1.0* (–)	.03	
Ope	.0* (–)	.16 (.05)	.08 (.05)	.01 (.05)	.03 (.05)	1.0* (–)	
Males							
Ext	–.29 (.06)	.93 (.09)	.11	.22	–.27	.28	
Agr	–.20 (.07)	.12 (.05)	1.37 (.14)	.26	–.31	.18	
Con	–.16 (.07)	.22 (.05)	.30 (.06)	.99 (.10)	–.31	.13	
Neu	–.54 (.07)	–.28 (.05)	–.40 (.07)	–.34 (.06)	1.21 (.11)	–.12	
Ope	.04 (.06)	.28 (.05)	.22 (.06)	.14 (.05)	–.13 (.05)	1.04 (.10)	

Note: Tabled values are parameter estimates, with standard errors in parentheses. Boldfaced parameter estimates had z-ratios greater than 5.0, italicized parameters had z-ratio between 2.0 and 5.0. In factor covariance matrices, covariances are shown below the diagonal, correlations above the diagonal.

* Asterisked values fixed at reported values to identify model.

Because strict factorial invariance held in Model 4, all between-group differences reside in estimates in the Ψ and α matrices, shown at the bottom of Table 20.5. The Ψ matrices contain covariances among the LVs; factor variances were fixed at 1.0 in the female sample and estimated in the male sample. Males appeared to differ in factor variance only on the Agreeableness factor, exhibiting greater variance, 1.37 (95% CI [1.14, 1.59]), than did females. Correlations among factors tended to be relatively low in each sample, ranging between .01 and .46 in absolute magnitude. Only 6 of the 10 interfactor correlations differed significantly from zero in the female sample, but all 10 of the correlations differed significantly from zero for males. To test whether variance-covariance differences across samples were significant, we fit a model that constrained the Ψ matrices to be invariant across samples. As shown in Table 20.4, this model, Model 5, fit the data significantly worse than did Model 4, $\Delta\chi^2(15) = 46.00, p < .0001$, but the practical fit indices were essentially unchanged. Thus, factor variances and covariances differed little across groups in a practical sense. Given the small worsening of practical fit for Model 5, some researchers might decide that this very highly restricted model is a better representation of the data than is Model 4, given the very parsimonious nature of Model 5.

Finally, mean differences between females and males are contained in the α vectors, shown at the bottom of Table 20.5. LV means were fixed at zero in the female sample, and mean differences for males were estimated. Because LV variances were fixed at 1.0 in the female sample, the mean differences on LVs for males were in a standardized, Cohen's d metric, indicating the number of LV SD units the males scored related to females. As seen in Table 20.5, significant but relatively small mean differences were found on the first three factors, with males reporting lower levels of Extraversion ($M = -0.29$), Agreeableness ($M = -0.20$), and Conscientiousness ($M = -0.16$) relative to females. The mean difference on the fourth LV, Neuroticism, was a more substantial, moderate-sized effect, with the male mean falling more than one-half SD lower than the

female mean ($M = -0.54$). The sex difference on Openness was nil. To test whether mean differences across samples were significant using an omnibus test, we fit a model that constrained the α matrices to be invariant across samples. This model, Model 6, fit the data much worse than did Model 4, $\Delta\chi^2(5) = 155.08$, $p < .0001$. Moreover, the practical fit indices were substantially worsened, and close fit based on the RMSEA was no longer tenable ($p < .005$). Thus, factor means differed significantly and to a practically important magnitude across groups.

Additional forms of invariance. The issue of invariance of factors across groups can be generalized across other dimensions explicit or implicit. For example, consider the dimension of time. Here the key scientific question concerns whether latent factors identified at one point in time or at one chronological age are the same as those at other points in time or age. This is a crucial question because tracking growth over time requires the implicit assumption that one is assessing the same construct across time or across the life span. Longitudinal CFA models can embody this assumption, especially if strong or strict factorial invariance holds across time. The testing of levels of invariance – from configural invariance to strict factorial invariance – follows the same steps for longitudinal data as outlined above for multiple-group modeling. The one difference in longitudinal models is the need to allow covariances among unique factors for the same indicator across the multiple times of measurement. This is a reasonable a priori specification, as unique factors are hypothesized to consist of a combination of specific (i.e., reliable) variance and random error. The longitudinal stability of the specific portion of the unique factor is therefore the basis for such covariances among uniquenesses. Among the many sources that could be offered, Ferrer, Balluerka, and Widaman (2008), Hancock, Kuo, and Lawrence (2001), McArdle (1988, 2007), Meredith and Horn (2001), Tisak and Meredith (1989), and Widaman, Ferrer, and Conger (2010) provide accessible discussions of details.

Forms of invariance across other blatant, or explicit, groups can be pursued; given the multitude of ways of grouping participants (e.g., race, culture, country, sexual orientation), an exhaustive list is impossible to develop. Moreover, methodological advances have enabled the search for latent groups (Muthén & Muthén, 1998–2012). In such models, a single sample is analyzed, but the researcher hypothesizes two or more latent groups within the single sample, groups that differ on certain model parameters. Models of this type are often referred to as factor mixture models, because the overall sample is viewed as a mixture of participants from multiple latent groups. If multiple latent groups are

justified statistically, each person receives a weight indicating the probability that she or he is a member of each group. Concerns have been voiced with regard to whether latent group results are real or artifactual (e.g. Bauer & Curran, 2003), but the next decade is ripe for continued research CFA models to investigate factorial invariance across latent and blatant groupings of participants.

Item Response Theory

Next we turn our attention to item response theory (IRT) modeling, another method for investigating measurement invariance. IRT is concerned with understanding and modeling the response process at the level of individual items. That is, the focus is placed at the item level and modeling the response process for each item. This contrasts with Classical Test Theory, where the emphasis is placed on test scores, for which reliability information (e.g., coefficient alpha; test-retest; see John & Benet-Martínez, Chapter 18 in this volume) is available, and with CFA. By emphasizing the item, we can use IRT to understand better the response process for each separate item and to identify whether extraneous factors influence response processes for individual items. IRT modeling was initially applied to cognitive and achievement tests where item responses were scored in dichotomous fashion: incorrect = 0; correct = 1. However, IRT models have since been extended to include polytomous response processes that are common in psychological and other social science research, where polytomous items are answered on scales with more than two values (e.g., Likert items on a 1 = “strongly disagree” to 5 = “strongly agree” continuum).

The core of an IRT model is a mathematical model for the probability of responding in each response category. Sometimes item responses represent the selection of a response option that falls on an ordered scale, whereas other items reflect selection of a response option that falls on no identifiable scale. For example, response categories can represent (a) success or failure on an item; (b) the specific response to an item from an ordered-categorical rating scale, such as the Center for Epidemiologic Studies – Depression Inventory, where choices range from 0 = “Rarely or none of the time” to 3 = “All of the time”; or (c) the specific choice (A, B, C, or D) to a multiple-choice item. Thus, IRT focuses on modeling, respectively, the likelihood or probability of (a) responding correctly to a test item, (b) responding using each of the four ordinal options from each item on the CES-D, or (c) responding to each of the unordered options on a multiple-choice item. These three likelihoods were chosen to represent three

different type of response processes for which different types of item response models are appropriate. Correct/incorrect response represents a dichotomous response process, rating scales yield responses on ordered categorical scales, and multiple-choice responses represent choices of unordered categories. We discuss models that are appropriate for each type of response process, beginning with dichotomous models, moving to ordered categorical polytomous models, and finally on to unordered categorical polytomous models.

Dichotomous Item Response Models

Item response models for dichotomous response processes are categorized by functional form, dimensionality, and the number of parameters per item. The functional form is often logistic, although initial models were based on the normal ogive model. In this chapter we discuss only logistic models as they represent the most common form of IRT model, but we note that models have been fit using other function forms (e.g., normal ogive) that give virtually identical results. The number of major dimensions underlying item responses is often restricted to one, but multidimensional IRT models have been discussed in the literature (Reckase, 1977; Wirth & Edwards, 2007). Here we limit our discussion to unidimensional models. The number of item parameters in dichotomous IRT models can range from one to four, although models with one, two, or three parameters are often utilized in research and are discussed here.

One-parameter logistic model. The one-parameter logistic model (1PLM) can be written as

$$P(x_{ij} = 1|\theta_i, \alpha, \beta_j) = \frac{\exp(\alpha(\theta_i - \beta_j))}{1 + \exp(\alpha(\theta_i - \beta_j))} \quad (20.10)$$

where $P(x_{ij} = 1|\theta_i, \alpha, \beta_j)$ is the probability of individual i responding in the higher of the two categories (e.g., correct) for item j conditional on item and person parameters. Item parameters include α , the discrimination parameter, and β_j , the location parameter. The only person parameter is θ_i , the location or score of person i on the latent trait. Only a single discrimination parameter is present in the 1PLM; that is, α does not vary over items. Thus, items differ only in one parameter – the location parameter β_j . The discrimination parameter α indicates the strength of the association between the items and latent trait, and the location parameter β_j is the point on the latent trait where persons have a 50% chance of

responding in the higher of the two categories. The α parameter is akin to a factor loading, and the β_j parameter is like a measurement intercept. Thus, the 1PLM is analogous to a one-factor CFA model in which all items have equal loadings on the LV, but items differ in their intercepts or difficulties.

Figure 20.2 is a plot of item characteristic curves (ICCs) for three items. In the ICCs the x-axis reflects the level of the latent trait, θ_i , and the y-axis reflects the probability of scoring a 1 (e.g., correct response). The three ICCs plotted correspond to three different items that vary with respect to their location parameters ($\beta_j = 1.3, .5, 1.2$), but have the same discrimination parameter ($\alpha = .8$). The location parameter is the point on the curve at which an individual has a .5 probability of correctly responding to the item, and the discrimination parameter, also referred to as the slope, corresponds to how rapidly the curve moves from the lower asymptote (probability equals 0) to the upper asymptote (probability equals 1). From this figure we can see how the probability of correctly responding to a test item increases as a function of the participant's underlying ability. Inspection of the figure also shows that (a) the ICCs are parallel (thus will never cross) because they have a common discrimination parameter; (b) items with a lower location parameter appear further to the left, indicating the item is easier to answer correctly for all levels of the latent trait, and (c) when the latent trait value equals the item location parameter ($\theta_i - \beta_j = 0$), the ICCs indicate a probability of .5 (or a 50% chance) of responding correctly to the item.

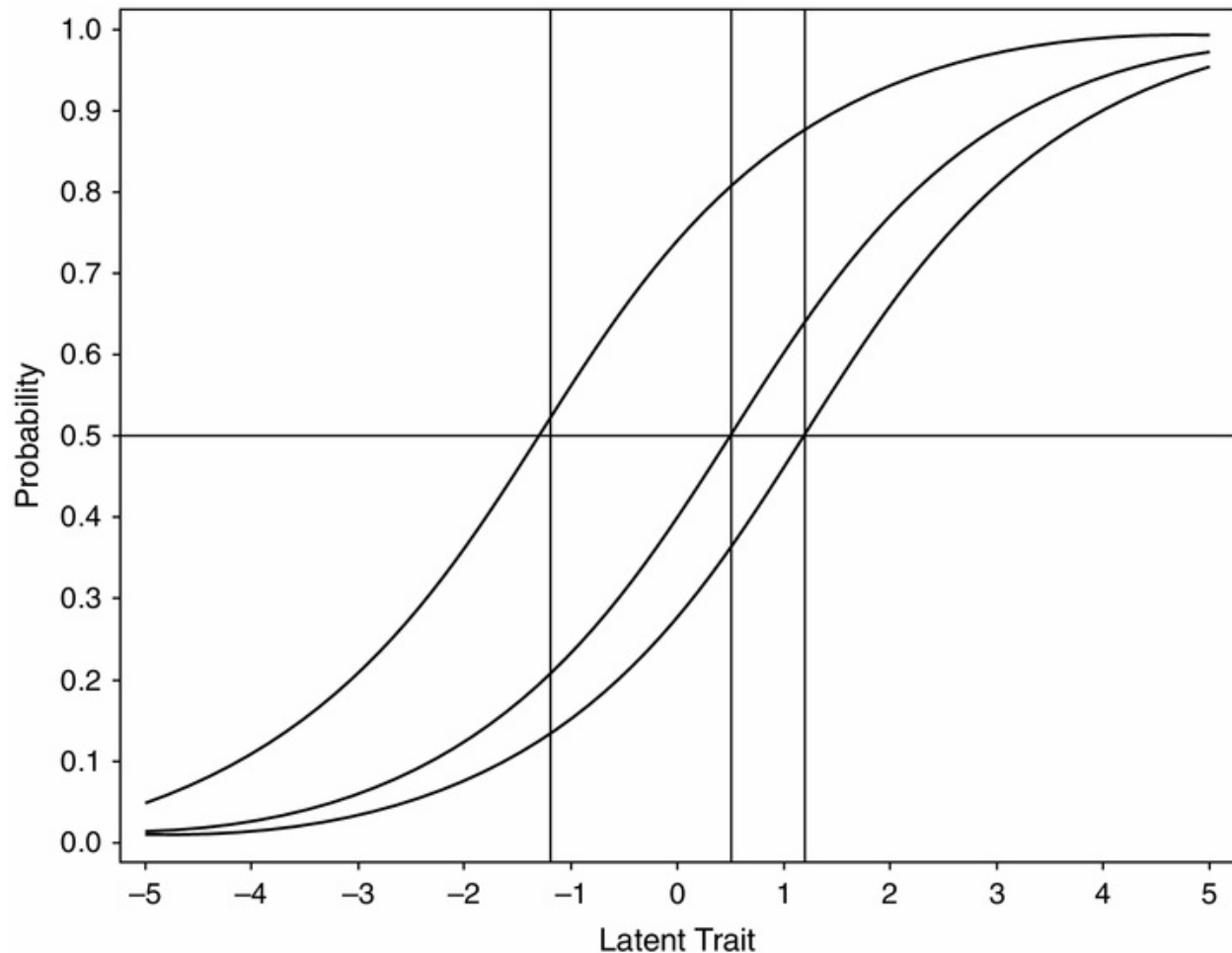


Figure 20.2. Item characteristic curves (ICCs) under a one-parameter logistic model (1PLM) for three items differing in difficulty.

The 1PLM is equivalent to the Rasch model (Rasch, 1960) and carries desirable measurement properties. These measurement properties include specific objectivity, the existence of a sufficient statistic, and objective measurement. Specific objectivity means that “the comparison of any two subjects can be carried out in such a way that no other parameters are involved than those of the two subjects, neither the parameter of any other subject nor any of the stimulus (i.e., measurement instrument) parameters” (Rasch, 1966, p. 92). In essence, person parameters are separable from the instrument parameters, a condition sometimes referred to as separability, parameter separation, or test-free measurement. Fundamentally, this property concerns measurement invariance under different tests or samples. That is, if a test is divided into two halves, with the first representing the easier items and the second representing the more difficult items, and estimation is carried out separately for each half, then

comparable latent trait estimates should be obtained indicating that the person parameters are not tethered to a specific test. The existence of a sufficient statistic means that the sum of the item scores (e.g., number of correct responses) has a one-to-one mapping with latent trait estimates; however, we note that this relationship is often nonlinear. Lastly, objective measurement is “the repetition of a unit amount that maintains its size, within an allowable range of error, no matter which instrument, intended to measure the variable of interest, is used and no matter who or what relevant person or thing is measured” (Program Committee of the Institute of Objective Measurement, December, 2000). Many experts have used objective measurement as justification for arguing that the 1PLM can provide interval-level measurement (assuming the data fit the model).

Two-parameter logistic model. The two-parameter logistic model (2PLM) can be written as:

$$P(x_{ij} = 1|\theta_i, \alpha_j, \beta_j) = \frac{\exp(\alpha_j(\theta_i - \beta_j))}{1 + \exp(\alpha_j(\theta_i - \beta_j))} \quad (20.11)$$

where $P(x_{ij} = 1|\theta_i, \alpha_j, \beta_j)$ is the probability that person i responds in the higher of the two categories for item j conditional on item and person parameters. Item parameters include α_j and β_j , the discrimination and location parameters, respectively. As in the 1PLM, the only person parameter is θ_i , person i 's latent trait score. The difference between the 2PLM and the 1PLM is that each item has its own discrimination parameter, indicating that items are allowed to differ in the strength of their association with the underlying latent trait. That is, items with larger discrimination parameters have a stronger association with the underlying latent trait – just as some items may load higher on a factor in a CFA.

To illustrate, [Figure 20.3](#) contains ICCs for two items that vary in both their discrimination and location parameters. The ICC for the first item shows a location parameter of -1 and a discrimination parameter of 1.2 , whereas the ICC for the second item has a location parameter of $+1$ and a discrimination parameter equal to $.4$. The first item has a larger discrimination parameter, evidenced in the figure by its steeper slope. The steeper slope indicates the item is better able to differentiate participants with similar latent trait scores. That is, a larger change in the predicted probability of correctly responding to the item occurs for participants with different latent trait scores near the location parameter. For example, for the highly discriminating item, a person with a latent trait score of -2 (1 logit below location) has approximately a $.23$

probability of answering the item correctly whereas a person with a latent trait score of 0 (1 logit above location) has approximately a .77 probability of correctly responding. This can be compared with the poorly discriminating item where a person with a latent trait score of 0 (1 logit below location) has approximately a .40 probability of answering the item correctly whereas a person with a latent trait score of 2 (1 logit above location) has approximately a .60 probability of correctly responding. Thus, an item with a larger discrimination parameter is better able to distinguish participants with higher latent trait scores from participants with lower latent trait scores. Larger discrimination parameters also reflect a closer association with the latent trait in much the same way that factor loadings in exploratory and confirmatory factor analysis reflect the degree of association between the indicator and the common factor.

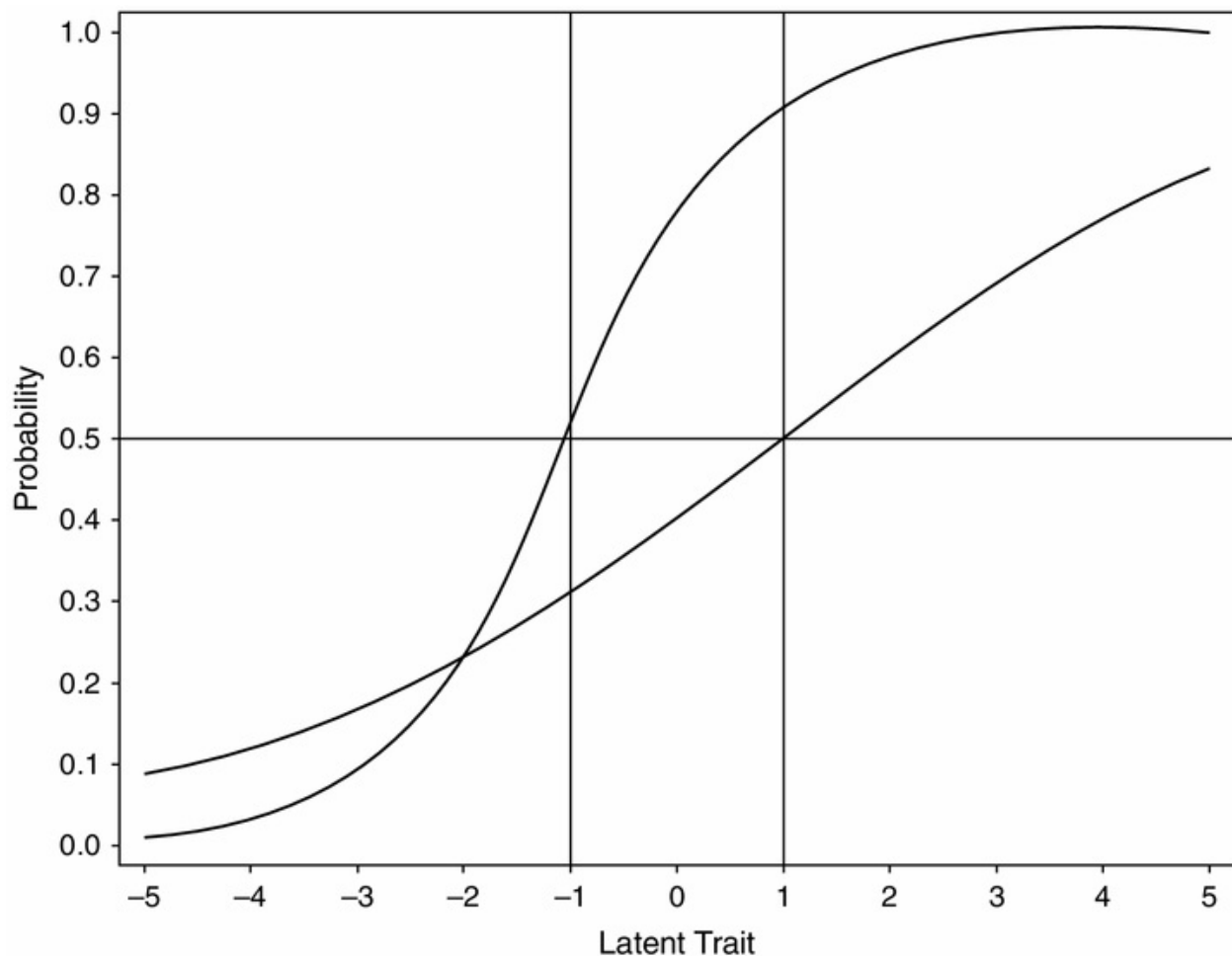


Figure 20.3. Item characteristic curves (ICCs) under a two-parameter logistic model (2PLM) for two items differing in difficulty and discrimination.

The 2PLM loses the desirable measurement properties associated with the

1PLM. However, the 2PLM often represents data better than the 1PLM as a result of its greater flexibility. The decision regarding the choice between the 1PLM and the 2PLM is often guided by statistical fit versus the desirable measurement properties of the 1PLM. Researchers can have strong opinions on which is more important, which often guides their decisions. That is, some experts feel that measurement properties of the 1PLM are fundamental, and data that do not conform to the 1PLM should be reevaluated (e.g., misfitting items removed). On the other hand, researchers may more heavily weigh the data collected than the measurement properties of the 1PLM. For these researchers, the 2PLM is often preferred.

Three-parameter logistic model. The three parameter logistic model (3PLM) can be written as

$$P(x_{ij} = 1 | \theta_i, \alpha_j, \beta_j, c_j) = c_j + (1 - c_j) \frac{\exp(\alpha_j(\theta_i - \beta_j))}{1 + \exp(\alpha_j(\theta_i - \beta_j))} \quad (20.12)$$

where $P(x_{ij} = 1 | \theta_i, \alpha_j, \beta_j, c_j)$ is the probability that individual i responds in the higher of the two categories for item j conditional on item and person parameters. Item parameters include α_j , β_j , and c_j , the item discrimination, location, and pseudo-guessing parameters, respectively. Once again, the only person parameter is θ_i , person i 's score on the latent trait. The difference between the 3PLM and the 2PLM is the pseudo-guessing parameter, which serves as a lower asymptote for the ICC. That is, no matter how low a person's latent trait, the probability of correctly responding equals c_j . This model is common for multiple-choice items to account for guessing the correct answer from the small number of alternatives.

Figure 20.4 contains ICCs for two items with different discrimination (1.5 and .7), location (−.8 and 1.1), and pseudo-guessing parameters (.3 and .1), respectively. The first item has a pseudo-guessing parameter equal to .3, leading to a lower asymptote at .30. The second item has a pseudo-guessing parameter of .1, so its ICC has a lower asymptote at .10, indicating that individuals have a lower probability of guessing the correct answer on the second item relative to the first item. As seen Figure 20.4, the location parameter is no longer the value of the latent trait where persons have a .50 probability of correctly responding because of the inclusion of the pseudo-guessing parameter; however, the

parameter is still related to the location or difficulty of the item.

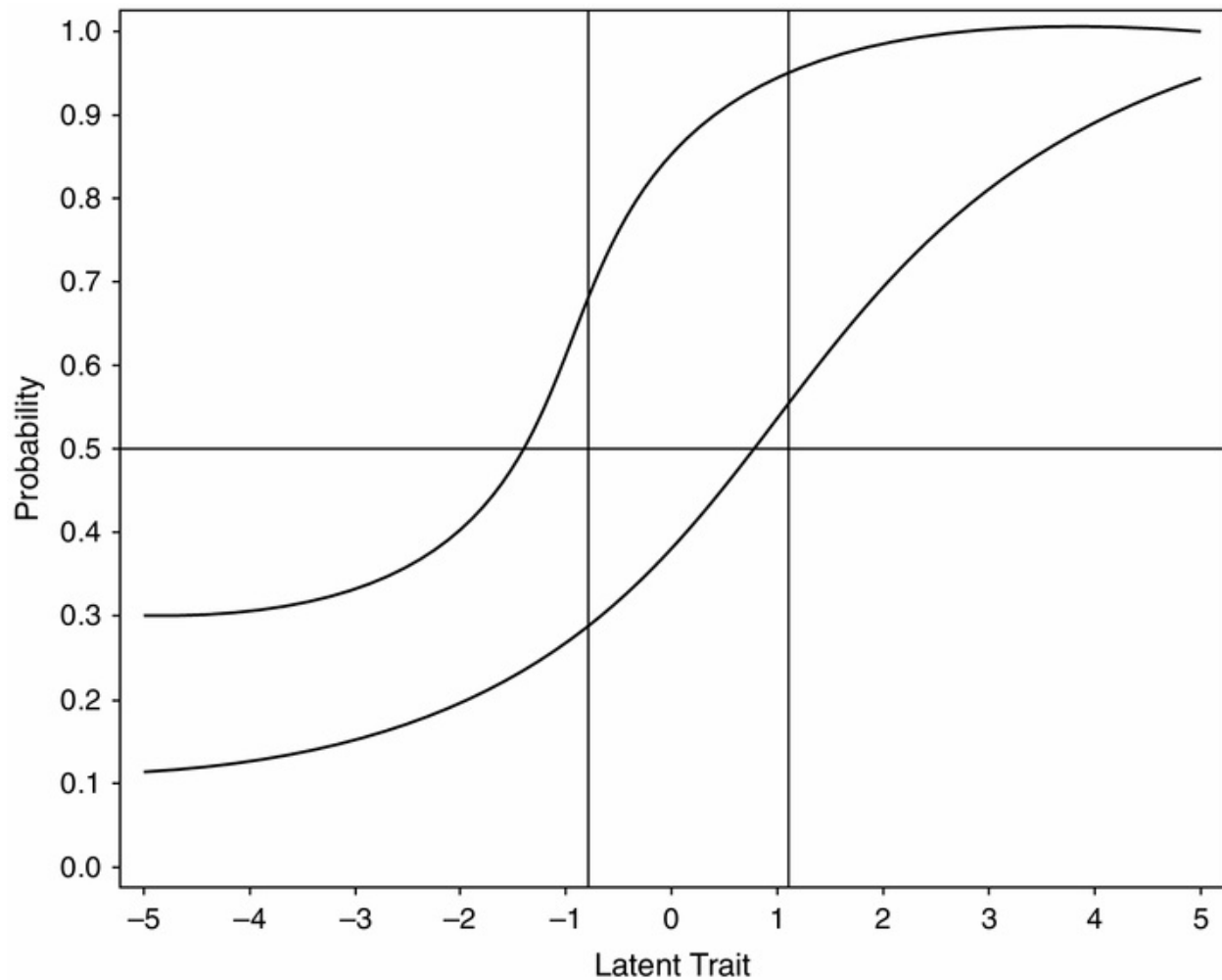


Figure 20.4. Item characteristic curves (ICCs) under a three-parameter logistic model (3PLM) for two items differing in difficulty, discrimination, and pseudo-guessing.

Polytomous Item Response Models

Several item response models are appropriate for polytomous response scales. In this discussion we fully describe two models and then discuss the remaining models as variations of these two models.

Graded response model. The first polytomous item response model is the Graded Response Model (GRM; Samejima, 1969), a straightforward extension of the 2PLM. The GRM can be written as

$$P^*(x_{ij} \geq c | \theta_i, \alpha_j, \beta_{jc}) = \frac{\exp(\alpha_j(\theta_i - \beta_{jc}))}{1 + \exp(\alpha_j(\theta_i - \beta_{jc}))} \quad (20.13)$$

where $P^*(x_{ij} \geq c | \theta_i, \alpha_j, \beta_{jc})$ is the probability of responding in or above category c conditional on person and item parameters. As with the 2PLM, item parameters include α_j , a discrimination parameter, and β_{jc} , the location parameter separating category $c - 1$ from c , and the person parameter is θ_i , the person's score on the latent trait. Figure 20.5 is a plot of Operating Characteristic Curves (OCCs) for a four-category item (0 = “never”, 1 = “sometimes”, 2 = “often”, 3 = “always”) with parameters $\alpha = 1.1$, $\beta_1 = -1.3$, $\beta_2 = -.1$, and $\beta_3 = 1.6$. The first curve represents the probability of responding in category 1 or higher; the second curve represents the probability of responding in category 2 or higher; and the third curve represents the probability of responding in category 3. From these curves, we see that it takes a larger difference in the underlying latent trait to go from category 2 (“often”) to category 3 (“always”) than the difference in the underlying latent trait to go from category 1 (“sometimes”) to 2 (“often”).

To calculate the probability of responding in each specific category, we assume that the probability of responding in category 0 – “never” or higher is 1 and subtract the probability of responding in or above category $c + 1$ from the probability of responding in or above category c . These category probabilities were calculated and are displayed in Figure 20.6. These curves are referred to as Response Characteristic Curves (RCCs). From this figure we can determine which category is most likely at each level of the underlying latent trait. We note that each response category is the most likely response at some value of the latent trait. Specifically, category 0 (“never”) is the most likely response for latent trait scores less than -0.9 ; category 1 (“sometimes”) is the most likely response for latent trait scores between -0.9 and -0.3 ; category 2 (“often”) is the most likely response for latent trait scores between -0.3 and $+1.4$; and category 3 (“always”) is the most likely response for latent trait scores greater than $+1.4$. OCCs and RCCs are important to inspect to understand how the item relates to the latent trait and how changes in the latent trait are reflected by changes in observed responses. These plots can be used to identify problematic items (e.g., items with categories that are never the most likely response).

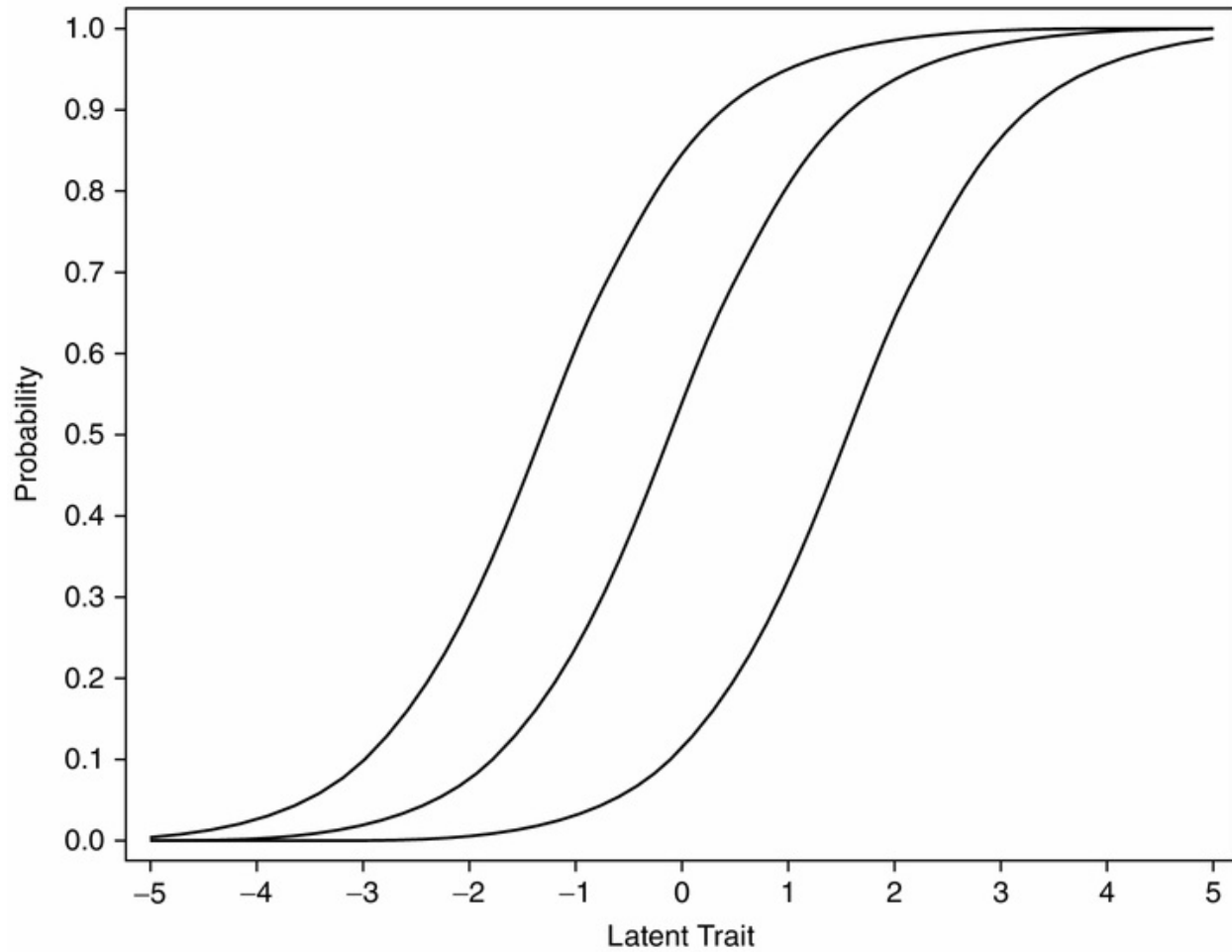


Figure 20.5. Operating characteristic curves (OCCs) for one item under the graded response model (GRM).

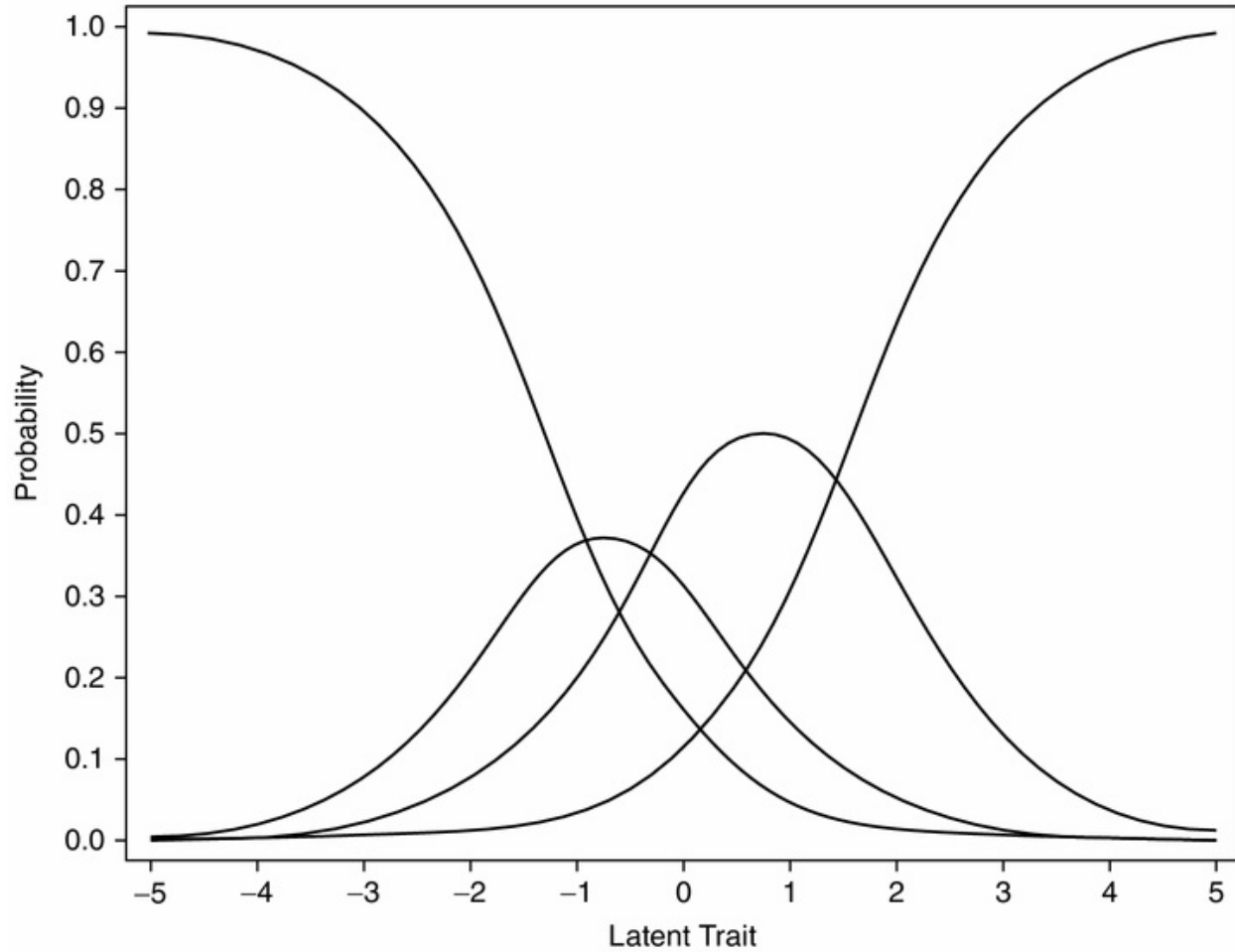


Figure 20.6. Response characteristic curves (RCCs) for one item under the graded response model (GRM).

Partial credit model. The second polytomous item response model is the Partial Credit Model (PCM; Masters, 1982) and is a polytomous extension of the 1PLM. That is, the PCM retains the measurement properties discussed for the 1PLM. The PCM can be written as

$$\begin{aligned}
 &P(X_{ij} = c | \theta_i, \alpha, \beta_{jc}, X_{ij} = c \text{ or } X_{ij} = c - 1) \\
 &= \frac{\exp(\alpha(\theta_i - \beta_{jc}))}{1 + \exp(\alpha(\theta_i - \beta_{jc}))}
 \end{aligned}
 \tag{20.14}$$

where $P(X_{ij} = c | \theta_i, \alpha, \beta_{jc}, X_{ij} = c \text{ or } X_{ij} = c - 1)$ is the probability of responding in category c conditional on person and item parameters and that the response was in category c or $c - 1$, α is the discrimination parameter, β_{jc} is the step difficulty

for separating response category $c - 1$ from c for item j , and θ_i is the latent trait score for person i . Probabilities from an item with the following PCM parameters: $\alpha = 1$, $\beta_{j1} = -1.6$, $\beta_{j2} = -0.6$, and $\beta_{j3} = 1.8$ were calculated and displayed in the RCCs in Figure 20.7. The interpretation of the step difficulties from the PCM are the locations where two RCCs cross indicating that responding in category c or category $c - 1$ are equal. For example, the first step difficulty is -1.6 ; at this theta value, the probability of responding in the first category and the probability of responding in the second category are equal and approximately equal to .42. We also note that the RCCs are not necessarily symmetric in the PCM as they are in the GRM.

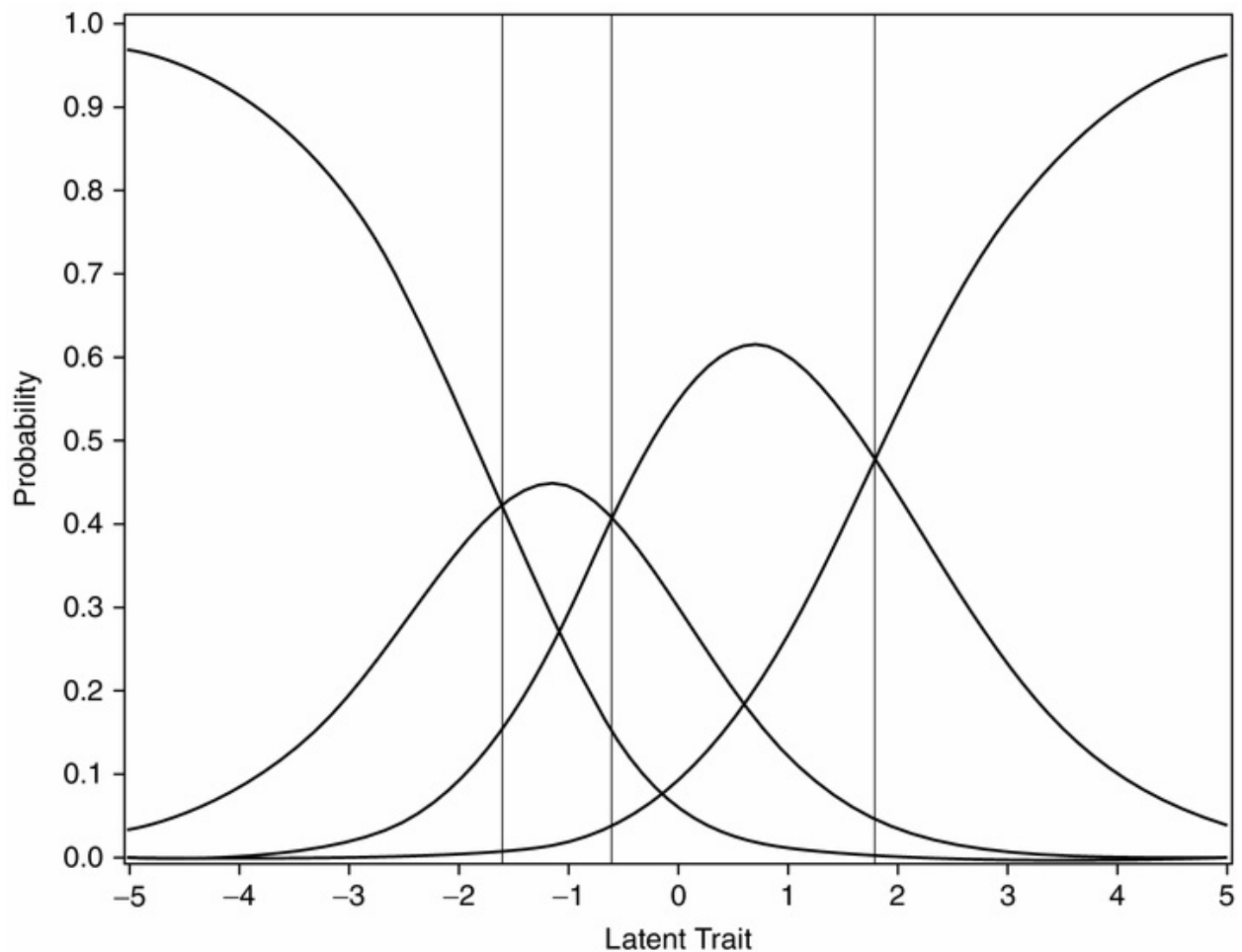


Figure 20.7. Response characteristic curves (RCCs) for one item under the partial credit model (PCM).

Other models. Additional polytomous response models impose or relax constraints in prior models. The Generalized Partial Credit Model (Muraki,

1993) is a PCM with a discrimination parameter for each item. The Rating Scale Model (Andrich, 1978) is a PCM with a constraint on the distance between the step difficulties across items. That is, $(\theta_i - \beta_{jc})$ is replaced with $(\theta_i - (\beta_j + \tau_c))$ where β_j is an item's location and τ_c is the step category for step c . So, the distances between adjacent response categories are the same, and items differ only with respect to location. Similarly, the Modified Graded Response Model (Muraki, 1990) imposes a similar constraint, but is based on the GRM.

The last polytomous response model we discuss is the Nominal Response Model (NRM; Bock, 1972). The NRM is appropriate for unordered responses and can be considered a generalization of Rasch-based polytomous models (Thissen & Steinberg, 1986). The NRM can be written as

$$P(X_{ij} = c | \theta_i, \alpha_{jc}, \beta_{jc}) = \frac{\exp(\alpha_{jc}\theta_i + \beta_{jc})}{\sum_{h=1}^{m_j} \exp(\alpha_{jh}\theta_i + \beta_{jh})} \quad (20.15)$$

where $P(X_{ij} = c | \theta_i, \alpha_{jc}, \beta_{jc})$ is the probability of responding in category c conditional on person and item parameters. Item parameters include α_{jc} , the slope parameter, and β_{jc} , the intercept parameter for category c of item j . As with other IRT models, the only person parameter is θ_i ; however, we note that each response category has its own discrimination parameter. The NRM compares the probability of responding in each category against a baseline category and can be referred to as a baseline logit model.

Item and Test Information

A commonly utilized statistic in item response modeling is item and test *information*. Item information is an index of the usefulness of an item for distinguishing among participants. That is, some items are not useful for distinguishing among participants whereas other items are highly useful. For example, imagine a simple addition question on a mathematics test that is given to high school students. In all likelihood, this item would not be useful in distinguishing among these students in terms of mathematical ability given the ease of the item relative to the students' mathematical ability. In the 2PLM, item information is calculated as

$$I_j(\theta) = \alpha_j^2 P(\theta)(1 - P(\theta)) \quad (20.16)$$

where $I_j(\theta)$ is item information for item j at a trait level of θ , α_j is item j 's discrimination parameter, and $P(\theta)$ is the probability of endorsing (responding correctly, scoring 1 versus 0) the item at a trait level of θ . This equation indicates that item information is greatest when the item is highly discriminating and the probability of endorsing the item is .5, which occurs when the latent trait and item difficulty are equal ($\theta_i = \beta_j$). Test information is simply the sum of item information curves under common IRT assumptions (unidimensionality and local independence),

$$I(\theta) = \sum_{j=1}^J I_j(\theta) \quad (20.17)$$

Test information is inversely related to the square of the standard error ($SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$), a measure of uncertainty in the location of the latent trait. We note that item information and therefore test information and the standard error are dependent on the value of the latent trait (θ). Therefore, tests can be seen as being reliable (having a small standard error) for certain individuals and unreliable for other individuals. For example, and extending our earlier example, a math test composed of several simple addition questions will have high test information (and therefore a small standard error) when the test is given to young elementary school students (appropriate target population); however, this test will have low information when administered to high school students. In most cases, it is optimal for test information to be relatively equal across the range of the latent trait for which the test is designed, indicating that participants are measured with the same level of precision. However, there are times where greater precision is needed for certain locations along the latent trait. One example is for certification tests where cut-off scores are identified to determine levels of proficiency. In these situations, test information should be greatest around the cut-off score to discriminate accurately between proficient and non-proficient test takers.

Data Requirements

In this section we consider appropriate sampling conditions and sample sizes when employing IRT models. An appropriate sample is one that spans the range of abilities/attitudes to be measured and one that is representative with respect to

the population selected. A sample that is uniformly distributed with regard to the construct of interest is ideal because good variation in response patterns will be seen across the range of abilities. This is especially important for multiple-category items because an appropriate (nonzero) number of responses should occur in each response category.

Regarding sample size, no good rules of thumb exist, because adequate sample size depends on many factors, including IRT model, number of items, population parameters, location of the items with respect to location of persons, and proposed use of the test. First, simpler item response models (e.g., Rasch-based models and dichotomous item response models) require fewer participants to obtain stable parameter estimates. Linacre (1994) suggested as few as 50 participants were adequate for the simplest Rasch models. With complex models (e.g., the 2PLM and especially the 3PLM), more participants are needed to obtain stable parameter estimates; however, how many more is not clear. Thissen, Steinberg, and Gerrard (1986) suggested as few as 200 participants can be adequate, whereas Tsutakawa and Johnson (1990) recommended 500 participants. A second factor related to sample size is the number of items, as a larger item pool requires a larger sample size to obtain stable parameter estimates. A third factor is the collection of item population parameters. That is, items with parameters that conform to the item response model require fewer participants to calibrate adequately. Reeve and Fayers (2005) noted that items that meet IRT assumptions of unidimensionality and local independence require fewer participants to calibrate well than items for which multidimensionality and local independence are an issue. Also, items that are poorly related to the underlying construct require larger samples (Thissen, 2003). A fourth consideration for sample size is the match or mismatch between locations of items and participants. Items that are well targeted for a sample (e.g., items where person ability and item difficulty are close) require fewer participants to obtain stable item parameter estimates. Orlando (2004) noted that a large homogeneous sample that does not reflect the population will result in highly precise parameter estimates for items within that narrow range of the underlying construct, but highly unstable parameter estimates for items outside this range (i.e., items that are not well targeted to the sample). Fifth, one must consider the desired use of estimates obtained from the item analysis. A smaller sample size is needed for a study to evaluate a scale's measurement properties compared to a study where the goal is to obtain precise IRT scores on which important decisions are based. In any case, simulation studies based on an appropriate population model are needed to determine adequate sample sizes.

Implementing IRT Models. Item response models can be fit in a variety of specialized stand-alone programs, such as Multilog (Thissen, 1991), Bilog (Zimowski, Muraki, Mislevy, & Bock, 1996), and WinSTEPS (Linacre & Wright, 1999). Additionally, item response models can be fit in statistical software packages, such as SAS (Sheu, Chen, Su, & Wang, 2005), R (Rizopoulos, 2006), and WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). Programs differ in the types of data and models that can be fit, the estimation routine (e.g., Maximum Likelihood, Marginal Maximum Likelihood, Joint Maximum Likelihood, and Bayesian Estimation), and the type of information reported in addition to parameter estimates (e.g., item fit statistics).

An appropriate starting point in any item analysis is a thorough investigation of dimensionality and local independence – two core assumptions when fitting *most* item response models. Dimensionality is best studied using item factor analysis, sometimes referred to as Full Information Factor Analysis and Nonlinear Factor Analysis (Wirth & Edwards, 2007; McDonald, 1967). Programs such as TestFact (Wilson, Wood, & Gibbons, 1991) and Mplus (Muthén & Muthén, 1998–2012) are able to conduct exploratory item factor analyses. Before fitting item response models, we often want to determine whether or not a dominant factor accounts for interitem associations. A variety of techniques can be used to determine the optimal number of factors, including the ratio of the first to the second eigenvalue, the number of eigenvalues greater than the average communality, parallel analysis, and a scree plot (Fabrigar & Wegener, Chapter 19 in this volume).

If a dominant factor is evident, then fitting the previously discussed IRT models is appropriate because these IRT models are unidimensional. If a dominant factor is not evident, then one may opt to fit IRT models to subsets of items that appear to represent identifiable constructs. Different researchers may take different approaches to model fitting. Some researchers who work solely with Rasch-based models place emphasis on the desirable measurement properties that accompany such models. After fitting a Rasch-based model, such researchers focus on item and person fit indicators, such as the INFIT and OUTFIT statistics available through WinSTEPS. Researchers may consider removing items from the scale if item fit statistics are poor. Again, we note that this approach emphasizes the measurement model, so the model is unchangeable, but the data can be changed.

Other researchers concentrate on non-Rasch models, acknowledging that Rasch models are often too restrictive (i.e., non-Rasch models tend to represent

item data better). In this approach, models can be compared based on overall model fit by comparing log-likelihood values. For example, the difference in -2 log-likelihood ($-2LL$) values is distributed as a chi-squared statistic with degrees of freedom equal to the difference in the number of estimated parameters between the two models. Comparing $-2LL$ values is an appropriate form of model comparison if the two models are nested, which means the first model is a restricted form of the second model. For example, the 1PLM is a restricted form of the 2PLM in that the discrimination parameter is equal across items in the 1PLM and estimated for each item in the 2PLM.

Regardless of the model-fitting strategy, a thorough study of local independence should follow the model fitting. Local dependence may manifest itself as minor factors that may not be evident in an exploratory factor analysis. Approaches to studying local dependence often involve an examination of item covariances after accounting for the underlying latent ability. For example, Yen (1984, 1993) proposed the Q3 statistic, based on the estimation of correlations among item residuals. Residual correlations greater than .2 are often interpreted as problematic (e.g., Chen & Thissen, 1997). Chen and Thissen (1997) proposed the G^2 statistic, a likelihood ratio statistic, which compares observed with expected responses from the fitted IRT model that assumes local independence. More recently, Edwards, Houts, and Cai (2012) proposed the Jackknifed Slope Index (JSI), in which item parameter estimates are obtained when all items are analyzed (full analysis) and, in subsequent steps, each item is removed one item at a time and revised item parameter estimates are obtained (reduced analysis). The JSI is the difference in the estimated discrimination parameter between the two analyses divided by the standard error of the discrimination parameter from the reduced analysis. If a discrimination parameter changes significantly with the removal of a single item, then the two items are considered to be locally dependent. Other indices are based on covariance structure models. For example, off-diagonal covariances can be examined from an exploratory factor analysis and modification indices can be examined from a confirmatory factor analysis (e.g., McDonald 1967, 1999; Steinberg & Thissen, 1996).

Empirical Example

To illustrate the use and application of item response models to psychological research, we analyzed items responses from the Agreeableness Scale from the Big 5 Personality Inventory. The data come from the 600 males and 600 females, ranging in age from 25 to 30 years ($M = 27.5$, $SD = 1.71$), who were the

basis for data in [Table 20.1](#). The Agreeableness Scale has nine items rated on a five point scale (1 = “Disagree Strongly”, 2 = “Disagree a little”, 3 = “Neither agree nor disagree”, 4 = “Agree a little”, 5 = “Agree Strongly”).

We began by fitting a nonlinear exploratory factor analysis model to the nine items from the Agreeableness Scale using Mplus (specifying items as categorical). The first four eigenvalues of the correlation matrix were 4.093, 1.140, 0.764, and 0.684, which suggests retaining one or two factors is reasonable. Standardized factor loadings from the one factor model ranged from .481 to .777, indicating moderate to strong associations with the underlying factor. Factor loadings for the two factor model indicated that items 2, 4, 5, 7, and 9 loaded on the first factor whereas items 1, 3, 6, and 8 loaded on the second factor. Examining the items from the Agreeableness scale, we noted that items 1, 3, 6, and 8 are negatively worded suggesting that the two factors may represent wording effects (i.e., positive versus negative wording). However, presence of a single dominant factor indicates that fitting IRT models is a reasonable next step.

Three IRT models were fit to the Agreeableness Scale. The three models were the Partial Credit Model, the Generalized Partial Credit Model, and the Graded Response Model. These models were fit using Multilog (v. 7.02). The Graded Response Model ($-2LL = 9,971$) had a lower log-likelihood than both the Generalized Partial Credit Model ($-2LL = 10,066$) and the Partial Credit Model ($-2LL = 10,248$), although the Graded Response Model cannot be directly compared to the Partial Credit Model or the Generalized Partial Credit Model because the models are not nested. Parameter estimates from the Graded Response Model are contained in [Table 20.6](#), and [Figure 20.8](#) contains the Response Characteristic Curves for the nine Agreeableness items. The RCCs indicate potential problems with certain items. For example, the RCCs for Item 6 shows that the third and fourth categories were never the most likely response. Thus, participants generally respond in the first, second, or fifth categories. Items 1, 3, 4, 5, and 8 showed similar problems to varying degrees. This contrasts with RCCs for items 2, 7, and 9, which showed adequate separation between response categories. Finally, the total information curve is plotted in [Figure 20.9](#). The total information curve shows the Agreeableness items provide greater information for participants at the lower end of the Agreeableness scale. Thus, this scale would benefit from having items that are more difficult to endorse positively in order to provide more information about participants at the upper end of the Agreeableness scale.

Table 20.6. Parameter Estimates from the Graded Response Model:

Agreeableness Scale from the BFI

Item Number	Discrimination α	Thresholds			
		β_1	β_2	β_3	β_4
1	1.10 (.08)	-1.63 (.14)	0.36 (.08)	1.07 (.10)	2.39 (.19)
2	1.66 (.11)	-3.24 (.24)	-2.26 (.24)	-1.32 (.08)	0.22 (.06)
3	1.17 (.09)	-3.59 (.31)	-1.60 (.13)	-0.91 (.09)	0.01 (.07)
4	1.46 (.10)	-2.46 (.17)	-1.38 (.10)	-0.83 (.07)	0.32 (.07)
5	1.29 (.09)	-3.11 (.14)	-1.92 (.14)	-1.16 (.10)	0.27 (.07)
6	.87 (.08)	-2.19 (.21)	-0.14 (.10)	0.59 (.11)	1.83 (.18)
7	2.39 (.14)	-2.70 (.17)	-1.73 (.09)	-1.06 (.06)	0.08 (.04)
8	1.42 (.09)	-2.16 (.15)	-0.39 (.07)	0.25 (.06)	1.12 (.09)
9	1.55 (.10)	-3.34 (.25)	-2.19 (.14)	-1.19 (.08)	0.41 (.06)

Note: N = 1,200 (600 females and 600 males in a single combined sample). Tabled values are parameter estimates, with SEs in parentheses.

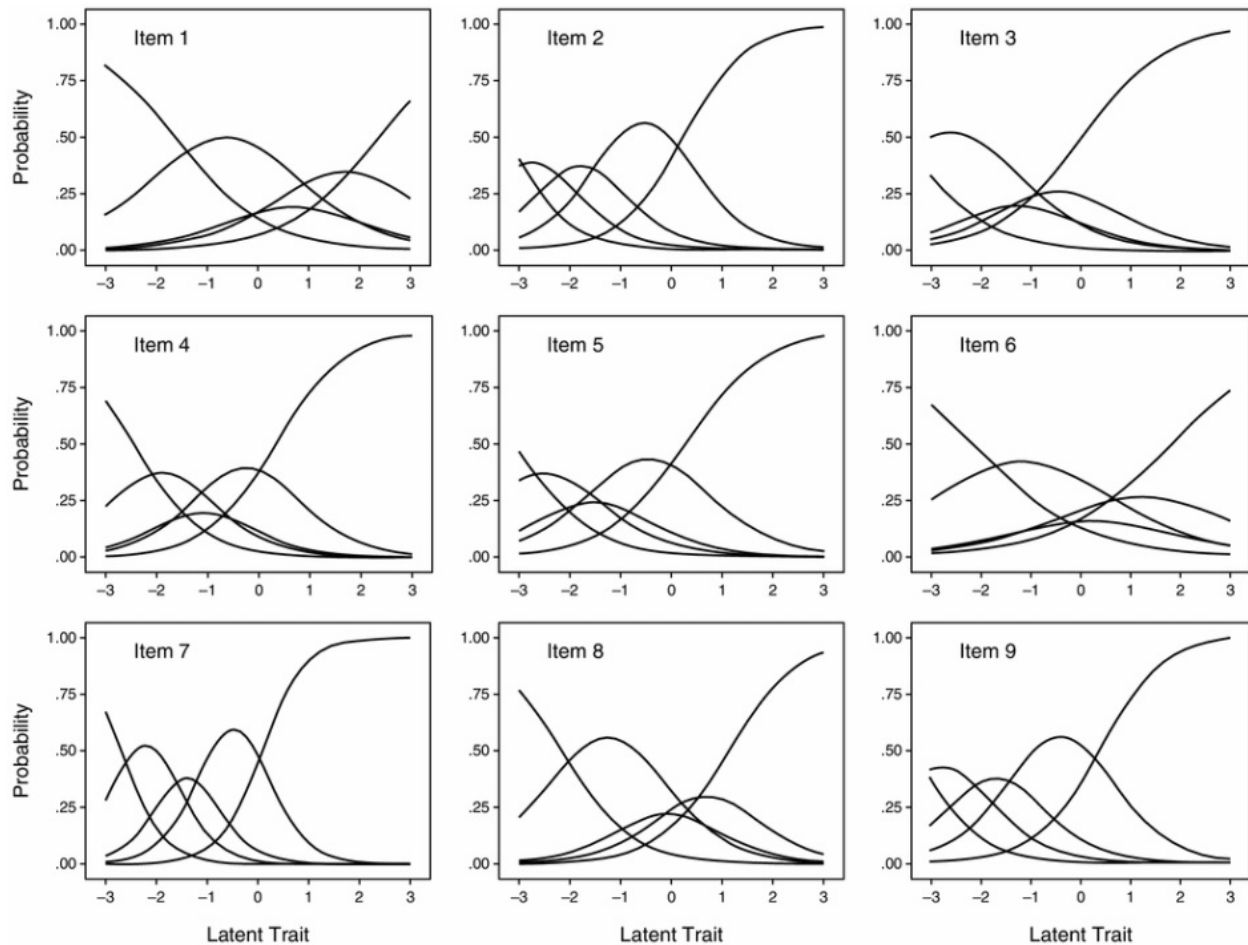


Figure 20.8. Response characteristic curves (RCCs) for the nine Agreeableness

items under the graded response model (GRM).

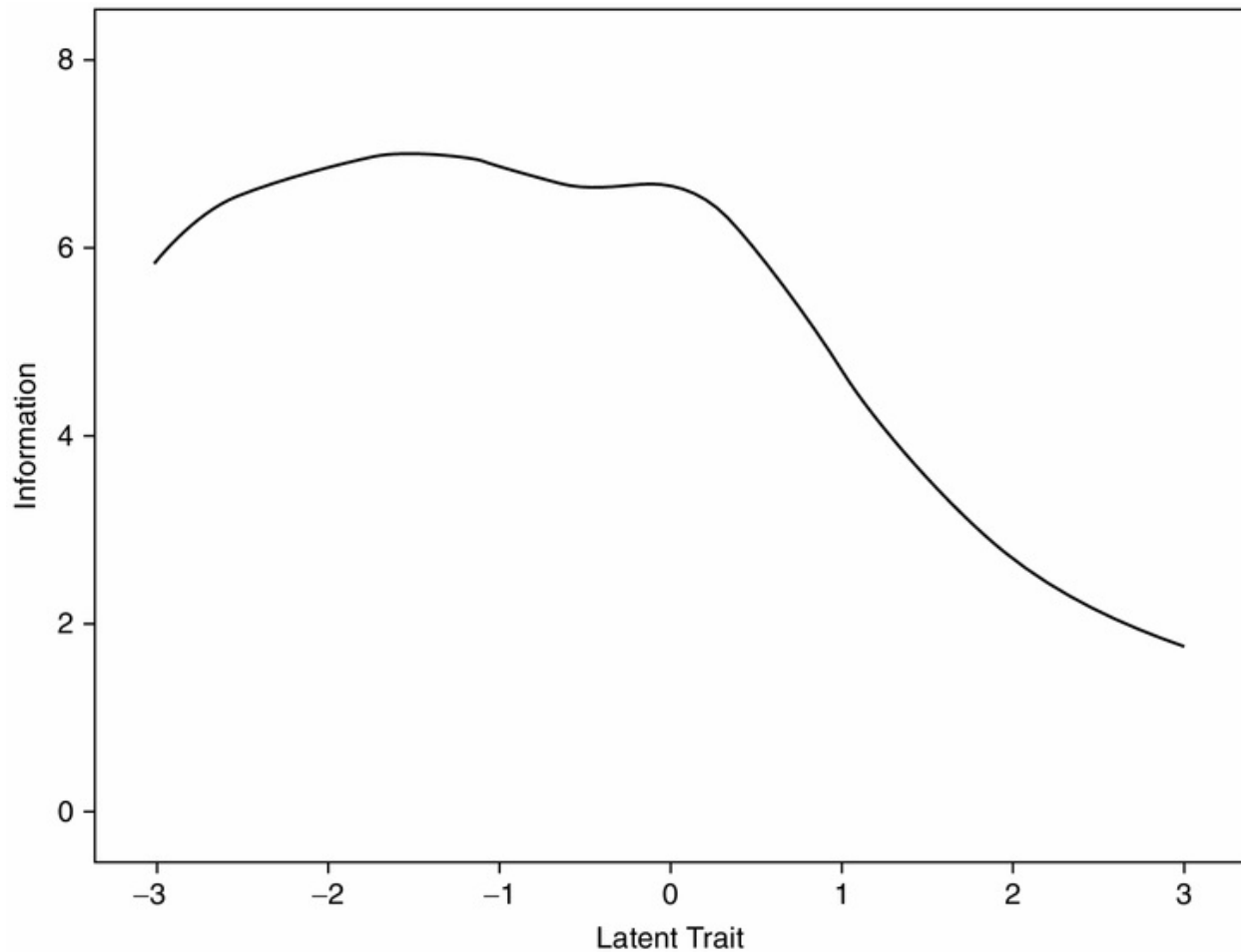


Figure 20.9. Plot of total scale information and estimated standard error of measurement (SEM) as a function of level of the latent trait for the Agreeableness scale, based on the graded response model (GRM).

We then examined local independence using IRTFIT (Bjorner, Smith, Stone, & Sun, 2007), which provides a statistical test of residual correlations. The statistical test is identified as a χ^2 statistic, which is distributed as a chi-squared variate with one degree of freedom. The IRTFIT program highlighted several potential local dependence issues with the Agreeableness Scale with multiple significant χ^2 values. However, an examination of residual correlations showed none was greater than .20, indicating the significant χ^2 statistics were strongly influenced by the large sample size.

Examining Measurement Invariance with IRT Models

General requirements. The study of measurement invariance using IRT is referred to as the study of Differential Item Functioning (DIF; Thissen, Steinberg, & Gerrard, 1986) and is focused on whether item parameters are invariant over measured groups, like gender, ethnicity, English language learner status, and so forth. Both the SEM and IRT frameworks are appropriate for examining item and test bias. However, SEM and IRT approaches to the study of measurement invariance often proceed in different ways. SEM tends to approach measurement invariance as an all-or-none decision – either measurement invariance holds or does not. In IRT, the study of measurement invariance proceeds item by item. In either framework, different analytic techniques can be used for studying measurement invariance (Woods, 2009). We describe the multiple-group method for IRT because of its common usage for studying measurement invariance and refer readers to Muthén (1985), Muthén and Lehman (1985), and Woods (2009) for a more in depth discussion of the MIMIC-Model based approach.

A major analytic approach to studying measurement invariance is the multiple-group framework where data are separated in two (or more) mutually exclusive groups and the IRT model is fit to the data for each group. One approach begins by fitting models for each group in which discrimination and location parameters are constrained to be equal for all items and the mean and variance of θ are separately estimated for each group. This is the “invariance” model because all item parameters are equated. Then, one by one, the discrimination and location parameters for a single item, referred to as the “target” item, are separately estimated for each group. The fit of this “alternate” model is compared with the “invariance” model. If the fit of the “alternate” does not differ significantly from the fit of the “invariance” model, then the process is repeated until all items are examined. If the difference in model fit is significant, this is noted because the target item *may* show bias and the process is repeated. The goal of this procedure is to determine the “anchor” items – items whose parameters are invariant over groups. This iterative process is continued until all items are examined. If the “alternate” model fits significantly better for multiple items, then the item that led to the largest improvement in fit is set aside, and the process is repeated until the “anchor” set is determined. Once the “anchor” set is determined, item parameters for all potentially biased items are re-examined for a lack of invariance with respect to the “anchor” set. If the “alternate” model fits significantly better than the “invariance” model, item parameters are examined separately to see if the discrimination and/or location parameters are the source of the change in model fit. If the location parameter is the only source of the

difference in model fit, then the item shows *uniform* DIF. If the discrimination parameter is a source of the difference in model fit, then the item shows *non-uniform* DIF. Uniform DIF indicates an item is universally easier (or harder depending on the direction of the group difference) for one of the groups. Non-uniform DIF indicates that the item is easier at certain levels of θ and harder for other levels.

Empirical example. Measurement invariance was examined for the nine Agreeableness items from the Big Five Inventory with respect to gender. In our illustrative data, the two groups consisted of 600 females and 600 males, and we utilized the multiple-group approach with iterative purification to determine the anchor items. The iterative purification procedure yielded six (of the nine) items to serve as the anchor set and identified three items as potentially having DIF (lack of measurement invariance). The invariance of the parameters from the three items was then evaluated, and the item parameters from the anchor set were constrained to be equal for males and females. We found that the three items showed uniform DIF – that is, the items differed only with respect to threshold parameters from the Graded Response Model, not discrimination parameters.

The Operating Characteristic Curves for these three items are contained in [Figures 20.10A](#) through [20.10C](#) for males and females separately. For Item 1, shown in [Figure 20.10A](#), the thresholds for males appear further to the right than the thresholds for females. Thus, on Item 1, males were likely to select lower categories for the same level of the latent trait compared to females, which is consistent with the mean difference on the Agreeableness LV in the CFA (cf. [Table 20.5](#)). On Item 3, shown in [Figure 20.10B](#), the major difference between males and females was the location of the first OCC threshold, which was shifted to the left for males. This shift indicated that males were more likely to respond above the first rating scale category than were females controlling for group differences in the latent trait. Remaining OCCs for Item 3 were rather similar across groups. Finally, OCC thresholds for item 5, shown in [Figure 20.10C](#), are strikingly different for males and females, with OCCs being much closer together for males than females. Thus, a smaller change in the underlying Agreeableness trait is needed for males to respond in a higher category compared to the amount of change needed for females.

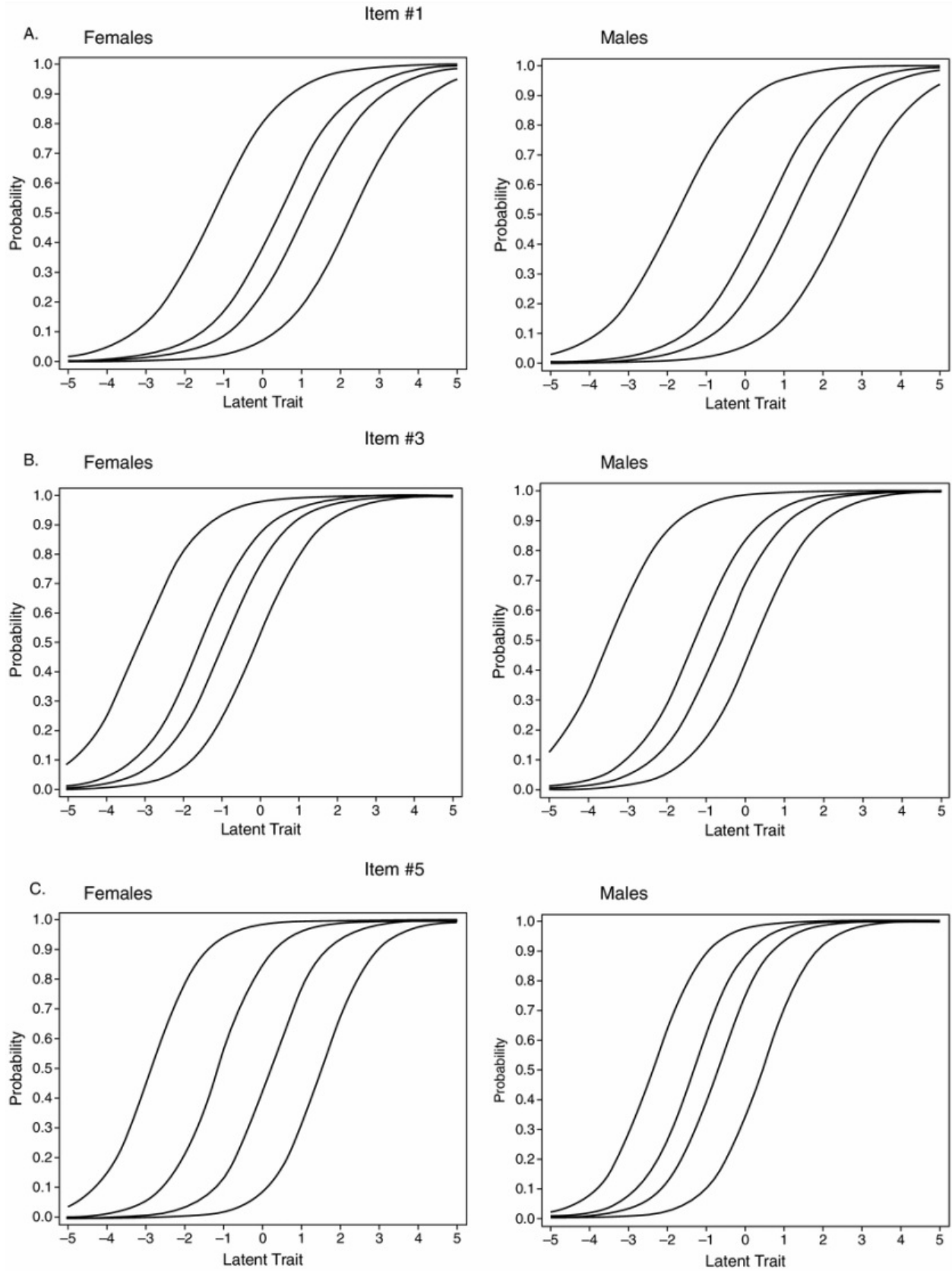


Figure 20.10. Response characteristic curves (RCCs) for the three Agreeableness items that exhibited differential item functioning under the graded response model (GRM). (A) Item 1. (B) Item 3. (C) Item 5.

Summary Comments

The first publication on factor analysis (Spearman, 1904) appeared more than a century ago, and the first paper on item response theory appeared more than 60 years ago (Lord, 1952). But, for many years, the use of these methods was the domain of the quantitative expert. Furthermore, well-formulated ideas about ways to use these techniques in a confirmatory fashion are relatively recent. CFA has been in existence for a little more than 40 years, and confirmatory model comparisons in IRT have been proposed during the past quarter century. Firm ideas about how to pursue measurement invariance using CFA and IRT models have become commonly accepted during just the past 15 years or so. Finally, relatively easy-to-use programs to perform all of these analyses are another recent development, and advances on this important front occur on a regular basis. The latter issue – the presence of easy-to-use programs to perform sophisticated analyses – is a very important one, as easy-to-use programs are needed to convert sophisticated analytic ideas from technical esoterica for the quantitative elite into techniques commonly used by practicing scientists.

The two techniques covered in this chapter, CFA and IRT, offer different but complementary ways of understanding psychological data and evaluating measurement invariance. Each method has its strengths, just as each has its weaknesses. But, measurement invariance is an important achievement in all research in the behavioral sciences, because invariance of our measuring instruments must hold for comparisons across groups to have any rational interpretation. We urge practicing scientists to wade into the methodology literature to begin to understand CFA and IRT, so that they can apply these methods in their research. Good introductory presentations are available for both CFA (e.g., Brown, 2006) and IRT (e.g., Embretson & Reise, 2000), and additional useful presentations are bound to appear in the near future. We look forward to increased use of CFA and IRT for investigating measurement invariance and improving the nature of measurements used in psychological sciences. Only by using measurements of the highest quality will our tests of important theoretical conjectures be based on the solid observational bedrock they deserve.

References

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, 8, 338–363.
- Bechtoldt, H. P. (1961). An empirical study of the factor analysis stability hypothesis. *Psychometrika*, 26, 405–432.
- Bechtoldt, H. P. (1974). A confirmatory analysis of the factor stability hypothesis. *Psychometrika*, 39, 319–326.
- Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin*, 107, 238–246.
- Bjorner, J. B., Smith, K. J., Stone, C., & Sun, X. (2007). *IRTFIT: A macro for item fit and local dependence tests under IRT models*. Retrieved November 3, 2011 from http://outcomes.cancer.gov/areas/measurement/irtfit_macro_users_guide.pdf.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29–51.
- Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Psychometrika*, 46, 443–459.
- Bollen, K. A. (1989). *Structural equations with latent variables*. Oxford: Wiley.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.
- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62–83.
- Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean and covariance structure models. In G. Arminger, C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185–249). New York: Plenum.

- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Cattell, R. B. (1956a). A shortened “basic English” version (Form C) of the 16 PF Questionnaire. *Journal of Social Psychology*, 44, 257–278.
- Cattell, R. B. (1956b). Validation and intensification of the Sixteen Personality Factor Questionnaire. *Journal of Clinical Psychology*, 12, 205–214.
- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54, 1–22.
- Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston: Houghton Mifflin.
- Cattell, R. B., & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement*, 24, 3–30.
- Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265–289.
- Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, 25, 1–27.
- Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology*, 100, 316–336.
- Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, 5, 115–124.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

- Edwards, M. C., Houts, C. R., & Cai, L. (2012). *A diagnostic procedure to detect departures from local independence in item response models*. Manuscript under review.
- Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, 65, 241–261.
- Eid, M., & Diener, E. (2006). *Handbook of multimethod measurement in psychology*. Washington, DC: American Psychological Association.
- Eid, M., Lischetzke, T., & Nussbeck, F. W. (2006). Structural equation models for multitrait-multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 283–299). Washington, DC: American Psychological Association.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. New York: Oxford University Press.
- Ferguson, G. A. (1941). The factorial interpretation of test difficulty. *Psychometrika*, 6, 323–329.
- Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order growth models. *Methodology*, 4, 22–36.
- Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology*, 44, 329–344.
- Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7, 286–299.
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48, 26–34.
- Goldberg, L. R., & Velicer, W. F. (2006). Principles of exploratory factor analysis. In S. Strack (Ed.), *Differentiating normal and abnormal personality* (2nd ed., pp. 209–237). New York: Springer.
- Gorsuch, R. L. (1988). Exploratory factor analysis. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed.,

- pp. 231–258). New York: Plenum.
- Guilford, J. P. (1941). The difficulty of a test and its factor composition. *Psychometrika*, 6, 67–77.
- Guilford, J. P. (1964). Zero correlations among tests of intellectual abilities. *Psychological Bulletin*, 61, 401–404.
- Hancock, G. R., Kuo, W.-L., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling*, 8, 470–489.
- Horn, J. L., & Hofer, S. M. (1992). Major abilities and development in the adult period. In R. J. Sternberg & C. A. Berg (Eds.), *Intellectual development* (pp. 44–99). New York: Cambridge University Press.
- Horn, J. L., & McArdle, J. J. (2007). Understanding human intelligence since Spearman. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 205–247). Mahwah, NJ: Erlbaum.
- Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist*, 1, 179–188.
- Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53–91). New York: Guilford.
- Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.
- John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory – Versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.
- Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32, 443–482.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183–202.

- Jöreskog, K. G. (1971a). Simultaneous factor analysis in several populations. *Psychometrika*, 36, 409–426.
- Jöreskog, K. G. (1971b). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36, 109–133.
- Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology*, 12, 247–252.
- Levy, R., & Hancock, G. R. (2007). A framework of statistical tests for comparing mean and covariance structure models. *Multivariate Behavioral Research*, 42, 33–66.
- Linacre, J. M. (1994). Sample size and item calibrations stability. *Rasch Measurement Transactions*, 7, 328.
- Linacre, J. M., & Wright, B. D. (1999). *A user's guide to Bigsteps/Winsteps*. Chicago: Mesa.
- Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, 32, 53–76.
- Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, 9, 151–173.
- Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with LVs: When “good” indicators are bad and “bad” indicators are good. *Psychological Methods*, 4, 192–211.
- Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, 18, 285–300.
- Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs*, No. 7.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.
- MacCallum, R. (1986). Specification searches in covariance structure modeling.

Psychological Bulletin, 100, 107–120.

- MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research*, 36, 611–637.
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84–99.
- Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, 13, 335–361.
- Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right – Camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, 18, 257–284.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselrode & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 561–614). New York: Plenum.
- McArdle, J. J. (1996). Current directions in structural factor analysis. *Current Directions in Psychological Science*, 5, 11–18.
- McArdle, J. J. (2007). Five steps in the structural factor analysis of longitudinal data. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 99–130). Mahwah, NJ: Erlbaum.
- McArdle, J. J., & Cattell, R. B. (1994). Structural equation models of factorial invariance in parallel proportional profiles and oblique confactor problems. *Multivariate Behavioral Research*, 29, 63–113.
- McCrae, R. R., & Costa, Jr., P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52, 81–90.
- McDonald, R. P. (1967). *Nonlinear factor analysis* (Psychometric Monograph

No. 15). Richmond, VA: Psychometric Corporation.

McDonald, R. P. (1999). *Test theory*. Mahwah, NJ: Erlbaum.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, 27, 82–99.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, 107, 247–255.

Meredith, W. (1964a). Notes on factorial invariance. *Psychometrika*, 29, 177–185.

Meredith, W. (1964b). Rotation to achieve factorial invariance. *Psychometrika*, 29, 187–206.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525–543.

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203–240). Washington, DC: American Psychological Association.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge/Taylor & Francis.

Millsap, R. E., & Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical perspectives and future directions* (pp. 131–152). Mahwah, NJ: Erlbaum.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, 39, 479–515.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, 14, 59–71.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, 17, 351–363.

Muthén, B. O. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43, 551–560.

Muthén, B. O. (1984). A general structural equation model with dichotomous,

- ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115–132.
- Muthén, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics*, 10, 121–132.
- Muthén, B. O., & Lehman, J. (1985). Multiple-group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics*, 10, 133–142.
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus User's Guide* (7th ed.) [Computer software]. Los Angeles, CA: Muthén & Muthén.
- Orlando, M. (2004). *Critical issues to address when applying item response theory (IRT) models*. Paper presented at the Drug Information Association meeting, Bethesda, MD.
- Program Committee of the Institute of Objective Measurement. (2000, December). *Definition of objective measurement*. Retrieved March 4, 2012 from <http://www.rasch.org/define.htm>.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Rasch, G. (1966). An item analysis that takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 19, 49–57.
- Reckase, M. D. (1977). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer-Verlag.
- Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. In P. Fayers & R. D. Hays (Eds.), *Assessing quality of life in clinical trials: Methods of practice* (pp. 55–73). New York: Oxford University Press.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analysis. *Journal of Statistical Software*, 17, 1–25.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of

graded scores. *Psychometric Monographs*, No. 17.

- Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & Clogg, C. C. (Eds.), *LVs analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.
- Savalei, V. (2010). Small sample statistics for incomplete nonnormal data: Extensions of complete data formulae and a Monte Carlo comparison. *Structural Equation Modeling*, 17, 241–264.
- Sheu, C.-F, Chen, C.-T., Su, Y.-H., & Wang, W.-C. (2005). Using SAS PROC NL MIXED to fit item response theory models. *Behavior Research Methods*, 37, 202–218.
- Sörbom, D. (1989). Model modification. *Psychometrika*, 54, 371–384.
- Spearman, C. (1904). “General intelligence,” objectively determined and measured. *American Journal of Psychology*, 15, 201–293.
- Spearman, C. (1927). *The abilities of man*. Oxford: Macmillan.
- Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, Iowa.
- Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods*, 1, 81–97.
- Sterba, S. K., & MacCallum, R. C. (2010). Variability in parameter estimates and model fit across repeated allocations of items to parcels. *Multivariate Behavioral Research*, 45, 322–358.
- Tellegen, A. (1982). *Brief manual for the Multidimensional Personality Questionnaire*. Unpublished manuscript, University of Minnesota, Minneapolis.
- Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory*. Chicago: Scientific Software.
- Thissen, D. (2003). Estimation in Multilog. In M. du Toit (Ed.), *IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact*. Lincolnwood, IL: Scientific Software

International.

- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567–577.
- Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, 99, 118–128.
- Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review*, 38, 406–427.
- Thurstone, L. L. (1934). The vectors of mind. *Psychological Review*, 41, 1–32.
- Thurstone, L. L. (1935). *The vectors of mind*. Chicago: University of Chicago Press.
- Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs*, No. 1. Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.
- Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs*, No. 2.
- Tisak, J., & Meredith, W. (1989). Exploratory longitudinal factor analysis in multiple populations. *Psychometrika*, 54, 261–281.
- Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of item uncertainty of parameter estimation on ability estimates. *Psychometrika*, 55, 371–390.
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–10.
- Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, 25, 1–28.
- Watson, D., Clark, L. A., Weber, K., Assenheimer, J. S., Strauss, M. E., & McCormick, R. A. (1995). Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. *Journal of Abnormal Psychology*, 104, 15–25.
- Watson, D., Weber, K., Assenheimer, J. S., Clark, L. A., Strauss, M. E., & McCormick, R. A. (1995). Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptom

- scales. *Journal of Abnormal Psychology*, 104, 3–14.
- Wherry, R. J., & Gaylord, R. H. (1944). Factor pattern of test items and tests as a function of the correlation coefficient: Content, difficulty, and constant error factors. *Psychometrika*, 9, 237–244.
- Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, 9, 1–26.
- Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, 28, 263–311.
- Widaman, K. F. (2007). Common factors versus components: Principals and principles, errors and misconceptions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 177–203). Mahwah, NJ: Erlbaum.
- Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, 4, 10–18.
- Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.
- Wilson, D. T., Wood, R., & Gibbons, R. D. (1991). *Testfact: Test scoring, item statistics, and item factor analysis*. Chicago: Scientific Software International.
- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12, 58–79.
- Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, 44, 1–27.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125–145.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing

local item dependence. *Journal of Educational Measurement*, 30, 187–213.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software.

* This research was supported by grant HD 064687 from the National Institute of Child Health and Human Development (Rand D. Conger, PI), and by grant DA 017902 from the National Institute on Drug Abuse and the National Institute on Alcohol Abuse and Alcoholism (Rand D. Conger, Richard W. Robins, and Keith F. Widaman, Joint PIs). Correspondence regarding this manuscript should be addressed to Keith F. Widaman, Department of Psychology, University of California, One Shields Avenue, Davis, CA 95616. E-mail address: kfwidaman@ucdavis.edu

Chapter twenty-one Multilevel and Longitudinal Modeling

Alexander M. Schoemann, Mijke Rhemtulla and Todd D. Little

Almost every statistical technique used in the social sciences makes the assumption that data are a simple random sample from a population of interest – that is, each member of a population has an equal chance of being selected. One implication of this assumption is that observations are *independent* – that is, participants' responses are not systematically more similar to each other than would be expected by chance. Frequently, of course, data are not so straightforwardly structured. Dependence can arise as a result of experimenter errors such as counting an observation twice or accidentally including a pair of siblings in a study, or more commonly, through a sampling design that includes a known dependence attributable to a nested structure.

Nested data are those in which observations are “nested” or “clustered” within groups in a hierarchical structure that is characterized, therefore, by “levels” of influence. For example, single measures from children are nested within classrooms, repeated measures are nested within individuals, or measures from mutual best friends or romantic couples are nested within dyads. Nested data may arise because of the inconvenience or even impossibility of collecting a simple random sample (e.g., trying to collect a large sample of children where none shared a classroom, a teacher, or a school would be prohibitive). Increasingly, however, researchers choose to collect nested data because of the opportunities it affords to examine effects at each level.

In this chapter we discuss methods for analyzing nested data with a focus on two statistical methods: multilevel modeling (MLM) and structural equation modeling (SEM). In the first part of this chapter we describe two *inappropriate* methods for analyzing nested data: aggregation and disaggregation. Next we cover the basic equations of MLM and discuss its uses with nested data. Finally we turn to the analysis of longitudinal data as a special case of nested data structures (i.e., where repeated measures are nested within individuals). We describe two types of SEM models for analyzing this sort of data, including panel models and latent growth curves. MLM and longitudinal SEM provide the

tools to not just correct for the independence violation in nested data but to systematically investigate relations between variables at two or more levels of analysis. In other words, the layered levels of influence can be disentangled and examined as key parameters of a statistical model. Theoretical models can be developed to understand and predict these key statistical parameters. For example, the big-fish-little-pond effect (Marsh et al., 2008) theorizes that at the individual level, academic achievement is positively related to academic self-concept (individual students with higher academic achievement also have higher academic self-concepts), but at the school level academic achievement and academic self-concept are negatively related (schools with high average academic achievement have lower average academic self-concepts). MLM and multilevel SEM allow researchers to examine these relationships in an accurate, parsimonious, and powerful manner.

Levels

It is common for multiple levels of nesting to exist; for example, McNulty and Russell (2010) investigated how problem-solving behavior affects newlywed couples' relationship satisfaction using a longitudinal design. In their first study, both members of a newlywed couple rated their relationship satisfaction every six months for a four-year period. Thus, in these data, observations (every six months) are nested within individuals (husbands and wives) and individuals are nested within dyads (newlywed couples). When discussing nested data, variables measured at the lowest level of the data hierarchy are referred to as level-1 variables, variables associated at the next higher cluster level are referred to as level-2 variables, and this naming process continues with additional levels of nesting. In the study by McNulty and Russell, variables measured every six months (e.g., relationship satisfaction) are level-1 variables, measures of individuals that are not measured over time, or that do not vary over time (e.g., gender), are level-2 variables, and measures concerning dyads (e.g., length of relationship) are level-3 variable s.

Aggregation and Disaggregation

When analyzing nested data, researchers have three types of options for analysis: aggregation, disaggregation, or analyses that take the nested structure of data into account (e.g., MLM or Multilevel SEM [MSEM]). We turn first to the potential pitfalls associated with aggregation and disaggregation (Table 21.1).

We then turn to the modern methods that allow one to embrace the dependencies in nested data structures and explicitly model the nested effects as theoretically meaningful and readily interpretable sources of influence.

Table 21.1. *Dangers of Aggregation and Disaggregation*

Technique	Statistical Danger	Interpretation Danger
Aggregation	Loss of Power	Ecological Fallacy
Disaggregation	Violation of Independence	Atomistic Fallacy

Aggregation is the process of averaging all level-1 units within their associated level-2 unit and performing analyses across all level-2 units. For example, in a study on work-team performance where data from individuals (level 1) are nested within teams (level 2), each workgroup might be assigned the mean of all individual scores, and the analysis would proceed on workgroups. *Disaggregation* is the process of analyzing all level-1 units and ignoring level-2 units. For example, in the same study on work-team performance, disaggregation would mean analyzing the individual-level data without accounting for the influence of team membership. Given that members of a team would influence each other, the influence of team membership would lead to biased results.

Aggregation and disaggregation have many downsides, both statistically and in terms of interpretation of results. A statistical danger of either method is the inability to detect relationships between variables that may only occur at a specific level. Recall that the relationship between academic performance and academic self-concept was positive at the student level (level 1) and negative the school level (level 2). If aggregation were used to analyze this relationship, researchers could not detect the level-1 relationship, and if disaggregation were used, researchers could not detect the level-2 relationship. Furthermore, aggregation or disaggregation can result in incorrect estimates of error variance, which in turn will lead to incorrect tests of effects of interest. In MLMs, error variance is modeled at both level 1 and level 2; aggregated data will have just one error variance, which represents some combination of the level-1 and level-2 errors. Assuming that there is, in fact, error variance at level 1 and level 2, this

can bias estimates of regression coefficients.

A statistical danger specific to aggregation is loss of power. When aggregating, the number of data points to be analyzed is the number of level-2 units, regardless of the number of level-1 units. For example, in a study of 50 workgroups with 20 individuals in each group, there are a total of 1,000 individuals in the study; aggregating across individuals would lead to analyses based on $N = 50$ level-2 workgroups. Disaggregation leads to the opposite problem. By making use of all observations and not accounting for their dependence, power is increased *too much* (assuming level-1 units within clusters are more similar than level-1 units between clusters, an assumption that almost always holds true in nested data). In other words, the standard errors of parameter estimates (e.g., regression coefficients) that we use to determine the significance of these estimates will be too small, leading to high type-I error rates.

To see how violations of independence affect type-I error rates, it is helpful to imagine an extreme case. Suppose a researcher intended to collect data from 100 students, but after collecting 25 observations, she then opted to copy those 25 rows of data three times to create 100 rows. Every observation is counted four times. Here it is obvious to see that by pretending to have $N = 100$ when really $N = 25$, power is artificially raised and type-I error rates will be high. Of course, taking several observations from within a workgroup is not as bad as copying one individual's data many times, but the idea is the same: Individuals within a workgroup share an environment (a socializing effect) as well as other attributes that led them to join the workgroup (a selection effect). Therefore, members of a given workgroup would be more similar to each other than two individuals chosen randomly from the population of interest. Two students from the same classroom share their learning environment, they interact with each other, and they may have been placed in the classroom because of some common attribute (e.g., neighborhood characteristics, learning disabilities, etc.). In such situations, children from the same classroom will, on average, be more similar to each other than will two students chosen from different classrooms. Two observations taken from the same person at different times will, on average, be more similar to each other than will two observations taken from different people. With nested data, the assumption of independence is almost *always* violated. In each case, treating observations as independent artificially increases power to the detriment of type-I error.

The interpretation dangers of aggregation and disaggregation are the

ecological and atomistic fallacy, respectively. The *ecological fallacy* is a logical fallacy where attributes of individuals are inferred from attributes of groups. This fallacy assumes that the characteristics of groups apply to all individuals within the group; however, there is actually more variability among individuals than among groups. For example, Robinson (1950) investigated the correlation between immigrant status and literacy. When he examined the relationship at the state level, the correlation was .46 (i.e., states with larger immigrant populations tended to have higher literacy). The ecological fallacy could lead someone interpreting this result to conclude that immigrants tend to be more literate than those born in the United States. When Robinson examined the relationship at the individual level, however, the correlation was only 0.11, suggesting a very different conclusion was appropriate for individuals (immigrant status is weakly related to literacy).

The *atomistic fallacy* is a logical fallacy where attributes of groups are inferred from attributes of individuals. This fallacy assumes that the relationships between variables measured at the individual level and relationships measured at the group level are identical. However, variables can have different meanings at individual and group levels and the relationship between variables can be very different when investigated at the individual or group levels. For example, imagine that a study finds that increases in income lead to increases in self-esteem. If, based on these data, we assume that countries with higher gross domestic products have higher self-esteem, we would be committing the atomistic fallacy.

A final danger related to inference is the problem of *contextual effects* (Lüdtke, Marsh, Robitzsch, Trautwein, Asparouhov, & Muthén, 2008), meaning that variables can have different meanings at different levels. In the workgroup example earlier in the chapter, the variable gender (contrast coded with male = -1 and female = 1) measured at level 1 (when using disaggregation) refers to the gender of each individual. However, when gender is measured at level 2, when using aggregation, the variable gender refers to the proportion of women in a given workgroup. At level 1, gender reflects the life-course socialization of gender and the biogenetic characteristics of being male or female. At level 2, gender is characteristic of the group environment and the social influences that different proportions of gender would carry. Thus, if one used disaggregation to analyze how gender (contrast coded with male = -1 and female = 1) predicts job satisfaction, the results would provide information about how men and women differ in their job satisfaction. However, if one used aggregation to analyze the same relationship, the results would provide information about how an

increasingly high proportion of females in a workgroup relates to job satisfaction.

In contrast to aggregation and disaggregation, multilevel modeling methods that explicitly account for the nested structure of data provide the best of all worlds. Statistically, these methods are as powerful as possible while also controlling type-I error. In terms of interpretation, these methods allow the researcher to estimate effects at both levels and interpret them separately. There are no meaningful downsides to multilevel modeling approaches, although they do require a brief period of learning to use correctly. In the next sections we discuss two types of analyses that can account for nesting: multilevel regression modeling (MLM) and multilevel structural equation modeling (SEM) .

Basics of Multilevel Modeling and Structural Equation Modeling

Both MLM and SEM can be used to analyze nested data. Since its inception, MLM has been used to analyze data that are nested as a result of cross-sectional or longitudinal designs. In contrast, SEM has primarily been used to analyze data that are nested as a result of longitudinal designs (repeated measures). Recent advances in software, however, allow SEM models to be fit to various nested data structures (this is called multilevel SEM, or MSEM). In fact, from a statistical perspective, MLM is simply a special case of MSEM. We begin with the general MLM case and then discuss various SEM models, particularly those related to longitudinal data structures .

Multilevel Modeling

MLM can be conceptualized as an extension of regression analysis that accounts for nested data and allows for effects to differ across level-2 units. The equation for a simple regression equation with one predictor is:

$$y_i = \beta_0 + \beta_1 X_i + e_i \quad (21.1)$$

where β_0 is the intercept, β_1 is the fixed slope for the entire sample, and e represents the residuals of the model which are assumed to be normally distributed with a mean of 0 and a variance of σ_e^2 . In this equation the subscript i

refers to a level-1 score (e.g., if level 1 is individuals, then y_i refers to individual i 's score on variable y). In MLM the intercept and slope can vary across level-2 units, and each intercept and slope can have its own distribution of residuals. For example, a researcher is interested in how gender predicts job satisfaction in a sample with individuals (level 1) nested within workgroup (level 2). Then x_i is the level-1 predictor gender (contrast coded with male = -1 and female = 1), and y is the level-1 dependent variable, job satisfaction. Each workgroup may have both a different average job satisfaction (as reflected in intercepts that vary across groups) and a different effect of gender on job satisfaction (as reflected in regression slopes that vary across groups). An equation for a two-level model with intercepts and slopes that differ across level-2 units would be:

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_{1j}X_{ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \tag{21.2}$$

where y_{ij} is a score for level 1 unit i nested within level 2 unit j , β_{0j} is the intercept or mean job satisfaction score for each of the j workgroups, β_{1j} is the slope for the effect of each i person's gender within each of the j workgroups (i.e., β_{1j} does not vary across individuals, but the predictor x_{ij} does), and e_{ij} is the person-level error within each workgroup. In other words, because the intercept (β_0) and slope (β_1) have j subscripts, they can vary over level-2 units; that is, the effect of gender (x) on job satisfaction (y) may be different in each workgroup. γ_{00} is the intercept or mean of the group means of all of the j workgroups' job satisfaction when all predictors are equal to 0. γ_{10} is the average slope for the effect of gender on job satisfaction across all workgroups. The residuals for the intercept (u_{0j}) and slope (u_{1j}) are assumed to be normally distributed with a mean of zero and variances of τ_{00} for the intercept and τ_{11} for the slope. τ_{00} is the variability in the intercept of job satisfaction across the j workgroups, and τ_{11} is the variability in the slope of gender on job satisfaction across the j workgroups.

Often the variance of the random effect is of substantive interest. For example, if τ_{11} is not equal to zero, then the slope of gender differs across workgroups. In other words, with significant variability in the random slope of gender, the difference between men and women in job satisfaction is not the same for all workgroups. When the intercept and slope vary across level-2 units, they are

known as *random effects*; intercepts and slopes that do not vary across level-2 units are known as *fixed effects*. Thus, Equation 21.2 contains a term for both a fixed effect of slope (γ_{10}) and a term for a random effect of slope (u_{1j}). Each of these effects can be tested. MLMs often contain fixed effects with no corresponding random effects; however, it is inadvisable to include a random slope or intercept if the corresponding fixed effect is not also included in the model.¹ If Equation 21.2 contained a random intercept but a fixed slope, then the equation would not contain the term u_{1j} , indicating that the slope of x does not differ across level-2 units. Through substitution, Equation 21.2 can be rewritten as:

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + u_{0j} + u_{1j}x_{ij} + e_{ij} \quad (21.3)$$

This equation can be divided into fixed effects (all portions of the equation containing γ) and random effects (the portions of the models with either e or u).

Within this basic format many different models can be estimated, including models with only intercepts, models with predictors at level 1, models with predictors at level 2, and models with predictors at both level 1 and level 2 (for examples, see Table 21.2). For example, a researcher may hypothesize that a workgroup's department (human resources or sales) also influences job satisfaction. Then w_j is a level-2 predictor (department, contrast coded as $-1 =$ human resources, $1 =$ sales), and y is a level-1 dependent variable such as job satisfaction (see Table 21.2 for the resulting equations). In this model, the level-2 predictor predicts the intercept of each workgroup (β_{0j}). A significant slope of the level-2 predictor (γ_{01}) indicates that job satisfaction is not the same for workgroups in human resources and sales. Finally, level-2 predictors can also predict level-1 slopes, resulting in a cross-level interaction between level-1 and level-2 variables. For example, the effect of gender on job satisfaction (a level-1 slope) may differ depending on the department housing a workgroup (a level-2 variable). Perhaps a researcher hypothesizes that men and women do not differ in their job satisfaction when their workgroup is in human resources but men and women differ in job satisfaction in workgroups in sales departments. In this model, the level-2 predictor (department) predicts both the random intercept (β_{0j}) and the random slope of gender (β_{1j}). In the reduced form of this model, there is now a slope (γ_{11}) for the interaction between gender (x_{ij}) and department (w_j) (see Table 21.2).

Table 21.2. Examples of MLMs

Model	Expanded Form	Reduced Form
Intercept only	$y_{ij} = \beta_{0j} + e_{ij}$ $\beta_{0j} = \gamma_{00} + u_{0j}$	$y_{ij} = \gamma_{00} + e_{ij} + u_{0j}$
Fixed level-1 predictor	$y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$ $\beta_{0j} = \gamma_{00} + u_{0j}$ $\beta_{1j} = \gamma_{10}$	$y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + e_{ij} + u_{0j}$
Random level-1 predictor	$y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$ $\beta_{0j} = \gamma_{00} + u_{0j}$ $\beta_{1j} = \gamma_{10} + u_{1j}$	$y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + e_{ij} + u_{0j} + u_{1j}X_{ij}$
Level-2 predictor*	$y_{ij} = \beta_{0j} + e_{ij}$ $\beta_{0j} = \gamma_{00} + \gamma_{01}w_j + u_{0j}$	$y_{ij} = \beta_{0j} + e_{ij}$
Level-2 and random level-1 predictors	$y_{ij} = \beta_{0j} + \beta_{1j}e_{ij}$ $\beta_{0j} = \gamma_{00} + \gamma_{01}w_j + u_{0j}$ $\beta_{1j} = \gamma_{10} + u_{1j}$	$y_{ij} = \gamma_{00} + \gamma_{01}w_j + \gamma_{10}e_{ij} + e_{ij} + u_{0j} + u_{1j}e_{ij}$

interaction between level-2	$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}$	$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}w_j + \gamma_{11}x_{ij}w_j + e_{ij} + u_{0j} + u_{1j}x_{ij}$
predictor and random	$\beta_{0j} = \gamma_{00} + \gamma_{11} + u_{0j}$	
level-1 predictor	$\beta_{1j} = \gamma_{10} + \gamma_{11} + u_{1j}$	

Note: All examples contain random intercepts.

* In a 2-level model, level-2 predictors are always fixed effects.

For models with more than two levels, random intercepts are estimated at all but the lowest level of nesting and random slopes can be estimated for variables measured at all but the highest level of the nesting structure. For a three-level model, random intercepts are estimated at both level 2 and level 3, the slopes of level-1 predictors can vary randomly across level-2 and level-3 units, and the slopes of level-2 predictors can vary randomly across level-3 units. For example, McNulty and Russell (2010) were interested in relationship satisfaction with three-level data (observations nested within person and persons nested within dyad). Thus, they could estimate random intercepts at both the person and dyad levels. The effect of a level-1 predictor such as negative problem solving behaviors (measured every six months) on relationship satisfaction could vary across individuals and across dyads, and the effect of a level-2 predictor such as gender could vary across dyads. For further information on multilevel modeling, researchers should consult Hox (2002), Raudenbush and Bryk (2002), and Snijders and Bosker (2011). For information on the application of multilevel modeling to dyads and groups, see Kenny and Kashy (Chapter 22 in this volume).

Intraclass Correlation Coefficient

An important consideration when using nested data analytic techniques is how much the nested structure of the data affects participants' responses on the

outcome variable. This cross-level influence can be quantified through the intraclass correlation (ICC), which is a ratio of the variance among level-2 units (τ_{00}) to the total variance of the outcome.

$$ICC = \frac{\tau_{00}}{\tau_{00} + \sigma_e^2} \quad (21.4)$$

The value of the ICC represents the proportion of variance in the outcome that results from differences among level-2 units, and values of the ICC range from zero to one. For example, in the workgroup example discussed earlier in the chapter, if the ICC of job satisfaction were 0.45, that would mean that 45% of the variance in job satisfaction stemmed from differences among workgroups and 55% of the variance in job satisfaction stemmed from differences among individuals. Small ICC values mean that a small proportion of variance is attributable to variation among level-2 units, and if the ICC is very small (e.g., less than .05, indicating that almost all of the variance in the dependent variable is attributable to individual-level variation), a multilevel approach may not be necessary. In this case, *disaggregation* may be appropriate. Large ICC values mean that a large proportion of variance stems from variation among level-2 units, and if the ICC is very large (e.g. greater than .95), again a multilevel approach may not be necessary. In this case, *aggregation* may be appropriate. The ICC can also be interpreted as the “expected average correlation between any two randomly chosen units that are within the same group” (Hox, 2002, p. 15). Small ICC values mean that units within a cluster are no more similar than units across different clusters; large ICC values mean that units within a cluster are much more similar to each other than to those in different clusters. If the ICC of job satisfaction was 0.45, then on average, levels of job satisfaction for two individuals in the same workgroup will correlate at $r = 0.45$. Each variable in an analysis will have its own ICC, reflecting the degree to which that variable exhibits dependency within clusters. Thus, job satisfaction might have an ICC of .45, but another dependent variable of interest, productivity judgments, may have an ICC of .66.

Structural Equation Modeling (SEM)

SEM is a multivariate data analysis technique that can be used to investigate relationships among both observed and latent variables. Latent variables are constructs that are measured by several observed indicators (e.g., *irascibility*

might be measured by a set of survey items such as “I get angry frequently” and “I find it easy to keep my cool”). Unlike observed scales, where item responses are typically summed to form a scale score, SEM allows for the common variance across the items to form a latent variable, while the unshared (i.e., unreliable) variance in each item is not included in the construct. In this sense, latent variables are like scale scores that are free of measurement error.

A full discussion of SEM is beyond the scope of this chapter. For readers who are entirely unfamiliar with SEM, very good introductions are available in Fabrigar and Wegener (Chapter 19 in this volume), as well as Brown (2006), Kline (2010), and Little (2013). For the remainder of this chapter we focus on how SEM can be used to model longitudinal data. SEM does not naturally account for nestedness in the way that MLM does, but it allows for complex models where nesting can be taken into account, especially when nesting stems from a longitudinal data structure. Growth curve models are a form of MLM in that observations are nested within individuals, and these models can be estimated using either MLM or SEM. Furthermore, Multilevel SEM (MSEM) blends MLM and SEM to account for group-nested data, but this extension of SEM is not yet widely available. With MSEM, researchers have the capability to estimate SEMs and account for non-longitudinal nested data structures such as persons nested in dyads or students nested in classrooms.

In addition to the advantages related to latent variables, SEM provides several advantages over MLM when dealing with longitudinal data. SEM can easily incorporate multiple constructs into the same model. This capability is particularly important when multiple dependent variables are of interest or when a variable acts as both a predictor and an outcome in the same model (e.g., panel models or mediation models). MLMs can only incorporate a single dependent variable, and variables in MLMs must function as either a predictor or an outcome, but not both. Furthermore, unlike MLM, SEM provides measures to evaluate how well a model fits the data. These fit measures allow researchers to determine if their model is correct given the data, or if some alternative model fits the data better (for more advantages of SEM, see Fabrigar and Wegener, Chapter 19 in this volume). In the following sections on panel designs and latent growth curves we explore in more depth how SEM handles nonindependence.

Common Research Designs with Nested Data

Research designs that result in nested data can be broken down into two broad

categories: cross-sectional designs and longitudinal designs. In a cross-sectional design, the nesting occurs at a single time point; for example, husbands and wives are nested within couples, individuals are nested within workgroups, or children are nested within classrooms. In a longitudinal design, the nesting occurs across time; for example, single observations are nested within individuals over time. These two types of designs are not mutually exclusive; it is possible to have a three-level structure where observations are nested within individuals, who are in turn nested within dyads (e.g., McNulty & Russell, 2010).

Cross-Sectional Designs

When nested data are cross-sectional, researchers can use MLM to evaluate how predictors, either at level 1 or level 2, affect the outcome variable. For example, Watson, Chemers, and Preiser (2001) investigated how basketball players' beliefs in collective efficacy were influenced by individual-level variables (e.g., optimism) and team-level variables (e.g., team size). An important limitation of MLM is that only level-1 variables may be used as outcome variables. Additional details on analyzing data from cross-sectional designs are addressed in Kenny & Kashy (Chapter 22 in this volume); the remainder of this chapter focuses on the analysis of longitudinal data.

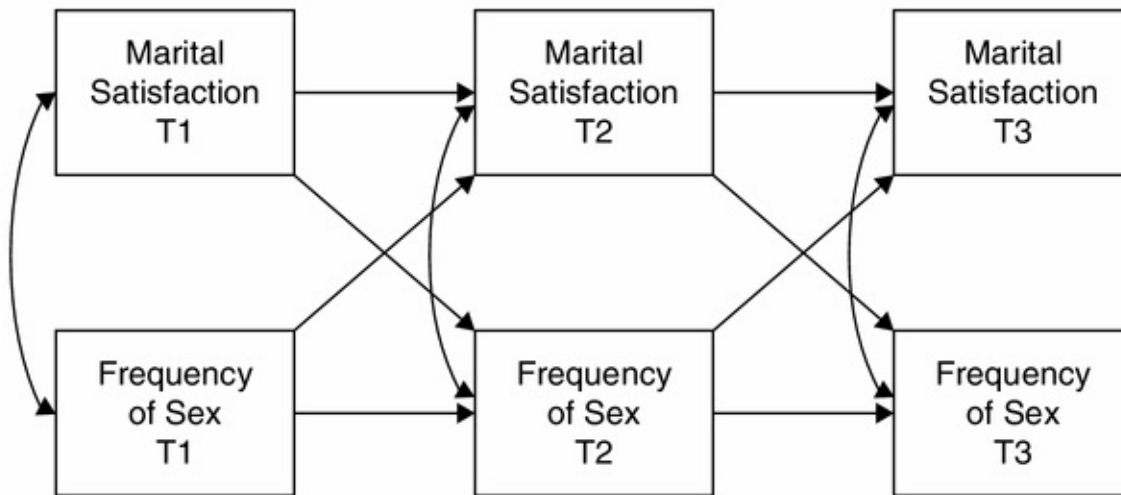
Longitudinal Designs

Before discussing traditional approaches to analyzing longitudinal data, we address a special case of longitudinal design that is common in personality and social psychology: repeated measures. In a repeated measures design, response (level 1) is nested within individual (level 2). Traditionally, data from a repeated measures design were analyzed using ANOVA; however, MLM provides a more accurate and flexible framework for analyzing repeated measures designs (Quené & van den Bergh, 2008). In MLM, a set of contrast codes, effect codes, or dummy codes (Wendorf, 2004) representing levels of the within-subjects factor (e.g., measurement occasions) are level-1 predictors. In the most straightforward case, with only level-1 predictors in the model, a single observation for each participant at each time point, and no missing data, MLM and ANOVA provide similar results. Both MLM and repeated-measures ANOVA/ANCOVA allow researchers to additionally include person-level predictors (i.e., level-2 predictors) and allow both continuous and categorical

level-2 predictors to interact with the within-subject (i.e., level-1) factor. However, MLM allows for level-1 slopes to vary randomly across level-2 units, which cannot be accomplished using ANOVA/ANCOVA. For example, Laurenceau, Barrett, and Rovine (2005) asked married couples (level 3) to complete a diary study over 42 days, and reported that daily levels of intimacy (an occasion-level, or level-1, variable) were predicted by both daily reported self-and partner-disclosure (occasion-level variables), as well as overall marital satisfaction (a person-level, or level-2, variable). In addition, MLM easily accommodates unbalanced designs and missing data.

Next we focus on two types of longitudinal analyses : panel designs and latent growth curves. Panel designs and latent growth curves differ in the questions they seek to answer: Panel designs investigate relationships among variables over time, whereas latent growth curves investigate the change in one variable over time. Panel designs can only be estimated using SEM or longitudinal regression. Either MLM or SEM can be used to analyze latent growth curves, and each type of analysis has advantages and limitations. MLM allows for many levels of nested data; however, it is generally limited to a single dependent variable, and it provides a relatively inflexible framework for analyzing longitudinal data. In contrast, SEM provides an extremely flexible framework for analyzing longitudinal data, allowing for multiple dependent variables and complex relationships among variables. However, SEM is limited to two (observation nested within individual) or, in the case of Multilevel SEM (MSEM), three levels of nested data.

(a) Panel Model with Observed Variables



(b) Panel Model with Latent Constructs

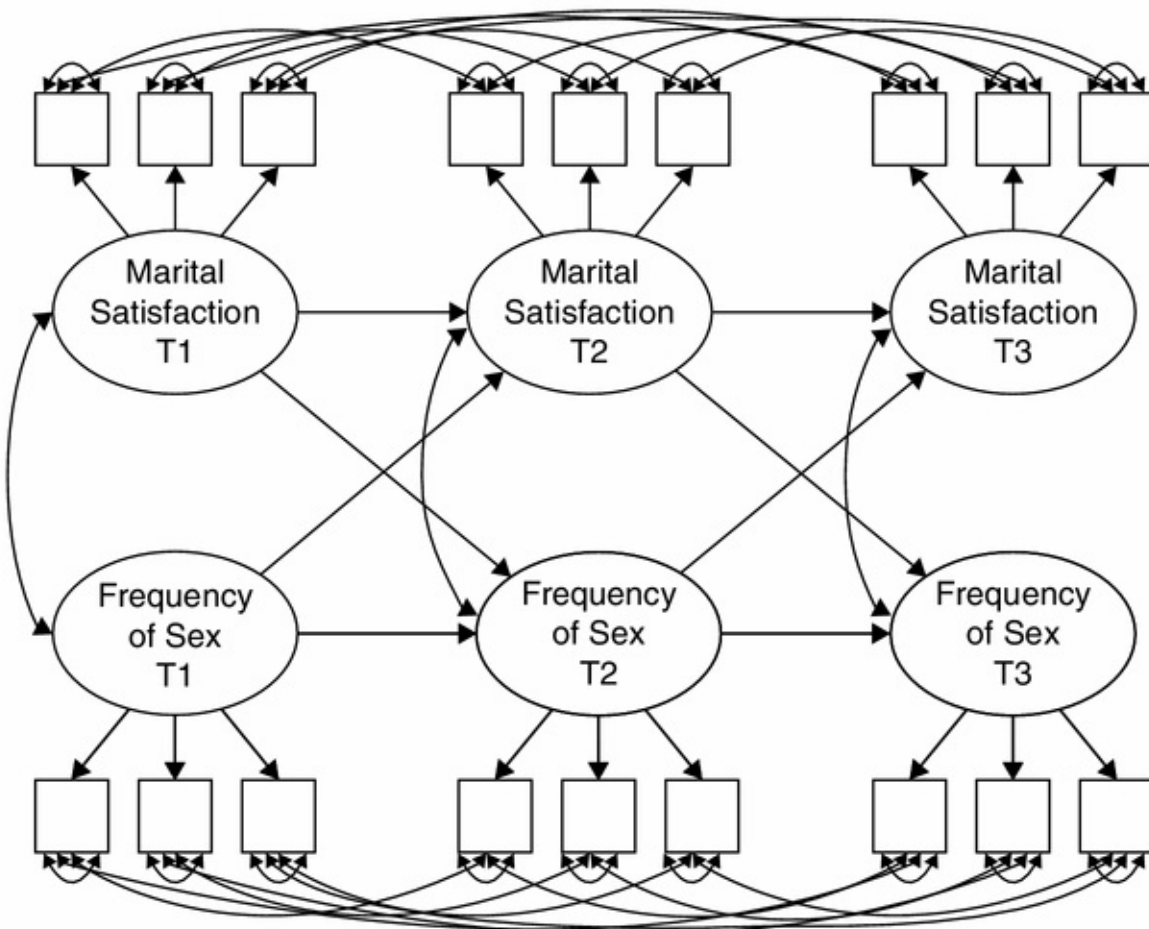


Figure 21.1. Panel models. The upper diagram (A) depicts a panel model where each construct is measured by a single variable at each of three time points. Regression paths between each variable and the variables at the previous

time point reveal the strength of predictive relations between these variables. The lower diagram (B) depicts the same model where each construct is measured by three observed indicators (e.g., survey items, subscales, observations). The latent variable model removes measurement error from the constructs, allowing the predictive relations between constructs to be estimated with much greater accuracy.

Panel Designs

Panel designs are appropriate to use when a number of constructs have been measured at several specific points in time from a sample of individuals. As in all longitudinal designs, repeated-measures data are nested within individuals, and all individuals are expected to have the same measurement occasions, missing data excepted. The primary goal of panel designs is to examine predictive relations between constructs across time. For example, one might be interested in how feelings of loneliness predict depression six months later, and in turn how levels of depression predict loneliness six months later. Or one might be interested in how frequency of sex in the first year of marriage predicts marital satisfaction several years later. One might further be interested in whether this relation is mediated by degree of sexual attraction to one's spouse, or whether it is moderated by having children in the interim. This focus is in contrast to latent growth models (next section) where the primary interest is in the growth or change of mean levels of constructs over time. In a panel design, all variables of interest are measured at several time points (at least two), so predictive relations at one or several time lags can be assessed while controlling for initial levels of each variable. The predictive regression relationships have a causal flavor to them but must be interpreted with caution. With many (if not most) panel models, statements about causality are not possible because the data are not controlled experimentally. Causality is implicated, however, to the degree that reliable predictive effects emerge over time. With panel models, the temporal separation between measurement occasions guides the direction of the predictive effects (predictive effects cannot go backwards in time, for example). In addition, paths that allow for control of the individuals' prior standing on a given construct are included (i.e., the autoregressive paths): These paths allow the predictive relations between constructs to be interpreted as partial relationships, controlling for prior levels of each outcome. While panel models can be constructed using only observed variables (e.g., a single variable to indicate sex frequency and another to indicate marital satisfaction), it is

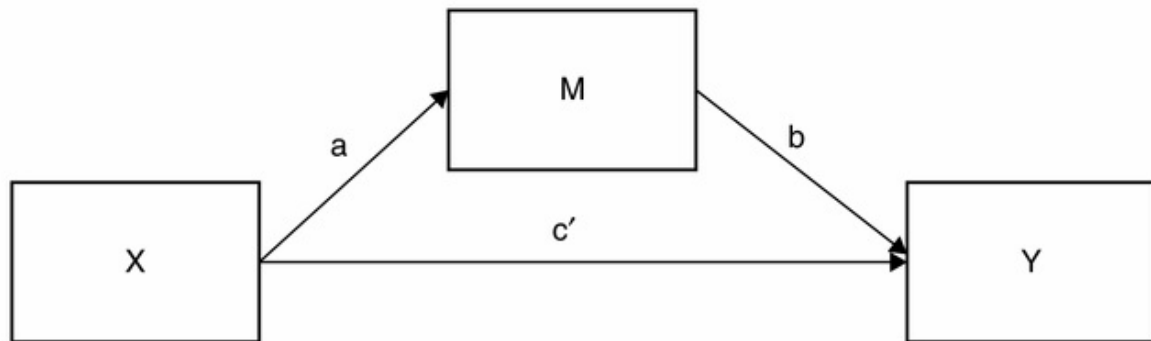
increasingly common and much more methodologically sound (both more valid and more reliable) to use latent variables with several observed indicators to access each construct of interest. Figure 21.1 depicts panel models with observed and latent constructs. When using latent variable models, however, it is necessary to ensure *factorial invariance* across time. Factorial invariance ensures that the latent constructs have the same definition across occasions; for example, that *marital satisfaction* is composed of the same indicators to the same degree at each occasion. Three increasingly strong levels of invariance are *configural invariance*, which tests whether each factor is measured by the same indicators at each time; *weak invariance*, which tests whether the relative contributions of each indicator's variance to the construct variance are the same at each time; and *strong invariance*, which tests whether the relative contributions of each indicator's mean to the construct mean are the same at each time. Of these three levels, only the first two are relevant to panel models, where we interpret the relations among factors (variances and covariances) but not their means. In latent growth curve analyses (see discussion later in the chapter), it is necessary to establish strong invariance to interpret the trends in factor means. We now explain each of these steps in slightly more detail (for a more thorough discussion, see Little, 2013).

Configural invariance. Configural invariance tests that each latent construct is made up of the same configuration of items at each time point. If *marital satisfaction* is indicated by three observed variables at the first time point, then it should be indicated by the same three observed variables (e.g., the same survey items or observations, assessed in the same way) at each subsequent time point. Testing configural invariance means testing the fit of a model where each latent construct is indicated by the same set of variables at each time point. Factor loadings, factor variances, residual variances, and intercepts are all freely estimated across time. Similarly, all latent factors are allowed to covary freely (i.e., only the measurement model is of interest; once invariance testing is complete, only then is structure imposed on the relations between latent factors). With the configural invariance model and every other invariance model, indicators' residual variances are allowed to covary with those of the same variable across time; for example, the residual variance of the first indicator of the first factor at each of three time points are allowed to freely covary.² If the configural invariance model fits reasonably well (evaluated using fit indices such as the chi-square test of exact fit, the RMSEA tests of close or not-close fit, or cut-off values for other fit indices), then configural invariance has been established and weak invariance can be tested.

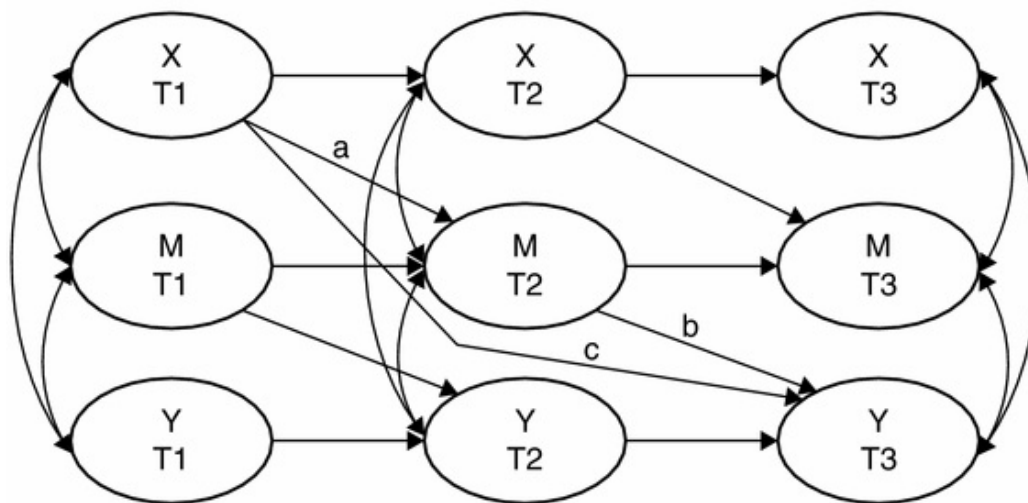
Weak invariance. Weak invariance means that each latent construct is made up of the same indicators *to the same degree*. In other words, the relative contribution of each observed indicator to each latent construct is the same at each time point. To test this, the set of factor loadings for each latent construct are set to be the same across time. Importantly, weak invariance does not assume that the latent variables have the same amount of total variance across time, so when the loadings are fixed to be equal across time, the variances at every time point subsequent to the first must be freed. To recap, weak invariance involves fixing factor loadings to be equal across time points, and freeing latent variances at every time point after the first (at the first time point, they are generally fixed to 1). Everything else about the configural model is unchanged: residuals are still allowed to covary across time, all latent variables are allowed to covary freely, and intercepts are freely estimated. Whether weak invariance holds can be tested by comparing this model to the configural invariant model. The weak invariant model imposes several constraints on the configural model, so it is nested within it. Because the chi-square difference test is overly sensitive to trivial differences, guidelines for determining whether invariance holds or not are based on simulation work that examines the change in relative fit statistics (e.g., Cheung & Rensvold, 2002; see also Little, 2013). If the model fit does not change substantially, weak invariance holds and the next step is to test strong invariance.

Strong invariance. Strong invariance tests whether the indicators of a construct contribute the same relative amount to the *means* of the latent variables. In other words, is the pattern of higher and lower indicator means the same across time? It is not necessary to test for strong invariance when using panel models, as only variances and covariances are relevant to these models. We describe strong invariance here in anticipation of our discussion of growth curve models later in the chapter.

(a) Cross Sectional Mediation Model



(b) Three Timepoint Mediation Model



(c) Two Timepoint "Half Longitudinal" Mediation Model

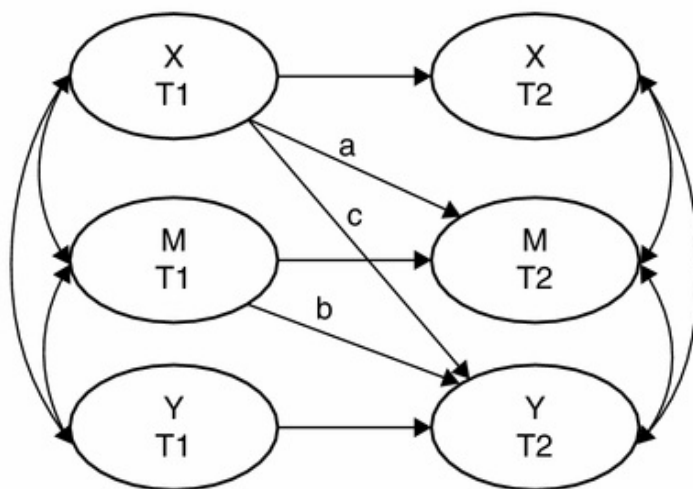


Figure 21.2. Mediation models. The upper diagram (A) depicts a simple cross-sectional mediation model. The middle diagram (B) depicts a three-time-point

longitudinal mediation model. The lower diagram (C) depicts a two-time-point longitudinal half-longitudinal mediation model.

The constraints imposed on the means when testing strong invariance are analogous to those imposed on the variances when testing weak invariance. The intercepts of a set of indicators of a factor are fixed to be the same across time. Just as we free the latent variances of all later time points when testing weak invariance, we also free the latent means of all later time points when testing strong invariance. Thus, the strong invariance model is identical to the weak model, but indicator intercepts are fixed to have the same values across time points, and latent means are fixed (typically to 0) at the first time point only, and freed at subsequent time points. As before, the strong invariant model is nested within the weak invariance model.

Little, Preacher, Selig, and Card (2007) discuss the steps and decision processes associated with invariance testing, including some guidelines on evaluating partial invariance when not all indicators of a construct are invariant. Little (2013) provides a number of didactic examples for testing longitudinal and multiple-group invariance. Because these issues are beyond the scope of this chapter, we will not discuss them further.

Testing predictive paths. Once weak invariance is established, it is possible to interpret the latent variables as having the same meaning across time. Now the structural model can be investigated. The covariances between all latent factors in the invariance-testing models are replaced with directional paths: *autoregressive* paths are those connecting a latent variable to itself at a subsequent time point, and *cross-lagged* paths are those connecting a latent variable to another variable at a subsequent time point. Within a time point, latent variables are allowed to covary freely (after the first time point, it is their residual variances that covary freely); directional paths within time are not estimated. Because previous levels of each construct are included in the model, each *cross-lagged* standardized path coefficient reflects the square root of the amount of variance in the construct's *change* since the previous time point that is predicted. For example, in Figure 21.1, the path from sex frequency at T2 to marital satisfaction at T3 is independent of marital satisfaction at T2, so the squared path coefficient can be interpreted as the proportion of change in marital satisfaction from T2 to T3 that is explained by sex frequency.

Testing mediation. If more than two variables are involved in the panel model, it is possible to test slightly more complicated hypotheses, such as

mediation. Mediation occurs when the effect of one variable (X) on another variable (Y), is transmitted through a “mediating” variable (M; see [Figure 21.2A](#)) (see Judd, Yzerbyt, & Muller, Chapter 25 in this volume). In this model the unconditional path from X to Y is known as the *c* pathway, the path from X to Y controlling for M is known as the *c'* pathway, the path from X to M is known as the *a* pathway, and the path from M to Y controlling for X is known as the *b* pathway (Baron & Kenny, 1986). The *indirect effect* (or mediation effect) is computed by multiplying *a* and *b*, and the significance of the indirect effect is assessed by testing if *ab* is different from 0. The indirect effect (*ab*) is not normally distributed and thus should be tested in a manner that does not assume normality. The most popular method of testing the indirect effect is through bootstrapping, which does not assume normality of *ab* (Shrout & Bolger, 2002). When bootstrapping, the researcher assumes that his or her sample data represent a population and takes thousands of samples, with replacement, from this “population” (for more information on bootstrapping in mediation, see Hayes, Preacher, & Myers, 2011; Rucker, Preacher, Tormala, & Petty, 2011; Shrout & Bolger, 2002). An analysis (i.e., the SEM panel model) is run on each bootstrapped sample, and the *ab* estimates are saved. These estimates form a distribution of *ab* and this distribution can be used to form a confidence interval or provide a significance test for *ab*.

Traditionally in social-personality psychology, mediation has been assessed in a cross-sectional design with variables X, M, and Y all measured at a single timepoint ([Figure 21.2A](#)). However, there is a potentially very serious problem related to testing mediation in a cross-sectional design. Mediation assumes a causal process: X causes changes in M, which in turn causes changes in Y. With cross-sectional data, the causal effects of X and M are impossible to assess because X, M, and Y are all measured at the same time. A significant indirect effect in a cross-sectional study may suggest (under extremely restrictive assumptions) a mediation effect, a spurious effect, or even a manipulation check (Fielder, Schott & Meiser, 2011). For example, a researcher may be interested in whether expressions of gratitude (being thanked) increase prosocial behavior, and whether this effect is mediated by feelings of self-worth (Grant & Gino, 2010). The researcher manipulates if participants are thanked or not (X) and measures self-worth (M) and prosocial behavior (Y) following the manipulation. The researcher can then test if being thanked causes increased self-worth (*a*) and that increased self-worth results in increased prosocial behavior (*b*) – the mediation hypothesis. However, an equally plausible model is one where being thanked causes increased prosocial behavior (*a*) and that increased prosocial

behavior results in increased self-worth (b). Furthermore, it is possible that a third, unmodeled variable, such as positive affect, is causing changes in both self-worth and prosocial behavior. In this cross-sectional design it is impossible to disentangle the causal relationship between M and Y, or to determine if changes in M and Y are being caused by some third, unmodeled variable.

By assessing mediation in a longitudinal panel model, many of the concerns about cross-sectional mediation can be alleviated. A longitudinal panel model with at least three time points (Figure 21.2B) allows for unbiased testing of mediation effects (Cole & Maxwell, 2003; Maxwell, Cole, & Mitchell, 2011). The longitudinal mediation model is a panel model with three variables, X, M, and Y, measured at three (or more) time points. The cross-lagged paths from X to M and M to Y represent the *a* and *b* pathways, respectively, and the significance of the indirect effect is still tested with *ab*. The temporal spacing of the longitudinal mediation model means that the *a* pathway represents the effect of X on the change in M since time 1 and the *b* pathway represents the effect of M on the change in Y since time 2. When using the longitudinal mediation model, it is important to consider the time between measurements. There must be enough time between measurement occasions for changes in M and Y to occur, otherwise it will be impossible to assess change. For an example, we will return to the example in which feelings of self-worth mediate the relationship between expressions of gratitude and prosocial behavior. Suppose that expressions of gratitude cause increased self-worth after five days and feelings of self-worth cause increased prosocial behavior after five days. If a researcher interested in this question measured expressions of gratitude, self-worth, and prosocial behavior three times, one day apart, he or she would likely conclude that self-worth is not a mediator. Conversely, if the researcher measured expressions of gratitude, self-worth, and prosocial behavior three times, five days apart, he or she would conclude that self-worth is a mediator of the effect of expressions of gratitude on prosocial behavior.

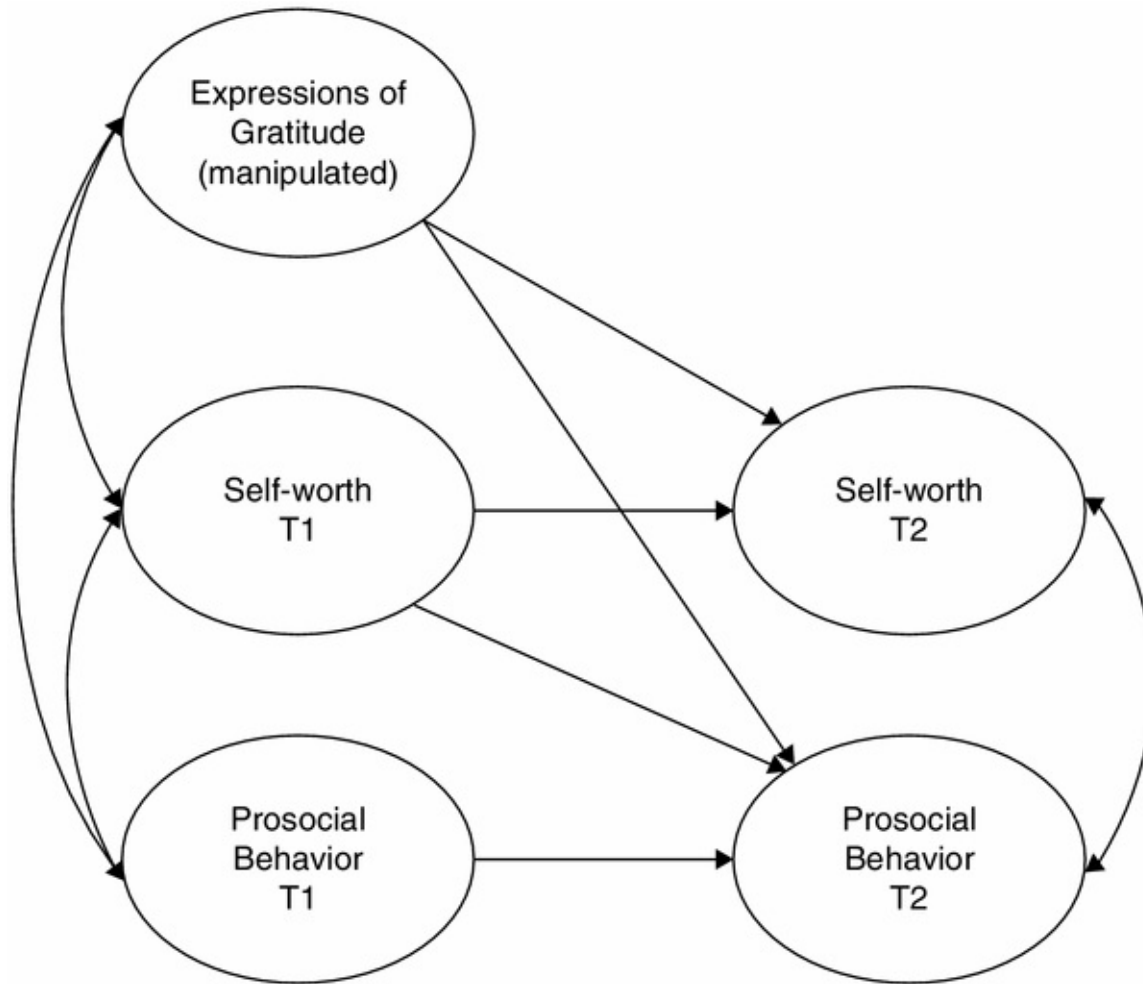


Figure 21.3. Example of a half-longitudinal mediation model with an experimental manipulation. In this model expressions of gratitude have been manipulated at time 1. Self-worth and prosocial behavior are measured before the manipulation (T1) and after the manipulation (T2).

Often in social-personality psychology it is not possible to measure variables across three time points, with sufficient time to observe changes in M and Y. In these situations, a *half-longitudinal* design (see [Figure 21.2C](#)) alleviates some of the concerns (Cole & Maxwell, 2003). The half-longitudinal design requires measuring X, M, and Y at two time points. The effect of X at time 1 on M at time 2 is the *a* pathway and the effect of M at time 1 on Y at time 2 is the *b* pathway. The half-longitudinal design allows researchers to investigate the effect of X and M on changes in M and Y, respectively, but it requires an important assumption: *stationarity*. Stationarity is the assumption that the causal relations between variables do not change over time. In the case of the half-longitudinal model, researchers must assume that the relationship between M at time 1 and Y

at time 2 does not change over time. If stationarity can be assumed, the a pathway represents the effect of X on the change in M since time 1, and the b pathway represents the effect of M on the change in Y since time 1. Further assumptions about directional dependency apply to infer causality (Pornprasertmanit & Little, 2013).

Either the longitudinal or half-longitudinal mediation model can be used when X is an experimental manipulation. If X is a manipulated variable, then it would be represented by a dummy (or contrast)-coded variable at time 1 and X would not appear at subsequent time points. For an example we will return to the study in which feelings of self-worth mediate the relationship between expressions of gratitude and prosocial behavior. In this example, expressions of gratitude have been manipulated so that participants are randomly assigned to either receive or not receive an expression of gratitude (Figure 21.3). Feelings of self-worth and prosocial behavior were measured before the manipulation and after the manipulation. Thus, the experiment uses a half-longitudinal mediation design to test mediation.

Latent Growth Curves

One of the most popular multilevel models for longitudinal data designs is the latent growth curve model (e.g., Preacher et al., 2008; Singer & Willett, 2003). Like panel models, growth curve models are multilevel models because observations are nested within individuals. In a panel model, the question being addressed is just about the level-2 associations. When a panel model is fit, the level-1 part of the model is a simple saturated model at each time point – that is, the scores for all individuals at each time point are characterized as a group mean and a variance; individual differences in rates of change are not modeled. Growth curve models, on the other hand, attempt to fit a model to the cross-time observations for each individual. These models typically include a latent intercept (i.e., a random intercept), which reflects the distribution of starting points before change has occurred, and a latent slope (i.e., a random slope of time), which reflects the distribution of rates of change over time. For example, we could fit a growth curve model to marital satisfaction measured on participants' wedding day and on each anniversary for five years. The latent intercept would reflect the distribution of marital satisfaction at the first time point (i.e., on the wedding day), and the latent slope would reflect the distribution of change in marital satisfaction over five years. Each of these parameters has a mean that can be meaningfully interpreted: The mean of the

intercept factor is the average marital satisfaction at the beginning of marriage, and the mean of the slope factor is the average rate of change in marital satisfaction over five years. The variances of each of these factors can also be meaningfully interpreted: The variance of the intercept is an estimate of the population variance in marital satisfaction at the beginning of marriage, and the slope variance is an estimate of the extent to which people differ in their rates of marital satisfaction change over time. In addition, the slope factor can take on different characteristics depending on the nature of change: It is most frequently characterized as a linear slope, but its shape can be freely estimated, exponential, or a linear slope can be supplemented with a second, quadratic slope.

The latent intercept and slope factors can, in turn, be characterized by a set of level-2 equations to examine the information contained in the mean and variances of the level-1 intercept and slope parameters. For example, we might be interested in looking at whether another variable such as age at marriage predicts the intercept or slope factors. If age is a significant predictor of the intercept, then we could conclude that age predicts marital satisfaction at the outset of marriage. If age is a significant predictor of the slope, then we could conclude that age predicts the rate at which marital satisfaction changes (increases or decreases) over the first five years of marriage. A more complex model could examine the intercepts and slopes of two variables simultaneously (e.g., marital satisfaction and frequency of sex) and look at whether the intercepts and slopes of both variables predict each other. Such a model might find that the slope of marital satisfaction is predicted by the intercept of sex frequency, suggesting that people who begin married life with high sex frequency have more rapidly increasing (or less rapidly decreasing) levels of marital satisfaction over five years.

These relationships can be seen clearly when represented using the standard multilevel regression framework shown in Equation 21.5. This equation displays a model for a growth curve model with a single predictor, X , that predicts both intercept and slope. This equation is perfectly parallel to the equations demonstrating a cross-level interaction in Table 21.2. In this equation an i subscript denotes individuals and an o subscript denotes measurement occasions. y_{io} is the dependent variable (e.g., marital satisfaction for a particular individual, i , at a particular occasion, o), β_{0i} is the latent intercept for individual i , β_{1i} is the latent slope for individual i , and ε_{io} is the leftover error for that particular individual at that particular occasion. The variable $time$ represents a coefficient for each measurement occasion; for example, in a linear growth curve, $time$

would equal 0 for the first measurement occasion, 1 for the second, 2 for the third, and so on. In turn, both the intercept and slope are modeled with their own equations because they vary across people. β_{0i} is composed of a mean (the average intercept across people), γ_{00} , and normally distributed residual variance, ζ_{0i} ; β_{1i} is composed of a mean (the average slope across people), γ_{01} , and normally distributed residual variance, ζ_{1i} ; Both the intercept and slope equations also include an optional level-2 predictor, X_1 (e.g., age at marriage). The parameters of the longitudinal model that is fit to each individual (as depicted by the i subscript) are used to represent changes over time (as depicted by the o subscript). The level-2 equations (i.e., the equations for β_{0j} and β_{1j}) are used to characterize these estimates at the level of the individual.

$$y_{io} = \beta_{0i} + \beta_{1i}(\text{time}) + \varepsilon_{io} \quad (21.5)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{10}X_1 + \zeta_{0j}$$

$$\beta_{1j} = \gamma_{01} + \gamma_{11}X_1 + \zeta_{1j}$$

Figure 21.4A shows what this model looks like in the SEM framework; MLM parameter labels are shown in parentheses. The single dependent variable in the MLM framework (y) is represented as a series of variables in SEM ($y_1 - y_4$), one for each measurement occasion (i.e., MLM uses long-format data where each row represents a measurement occasion within a participant whereas SEM uses wide-format data where each row represents a participant). The coefficients on the paths from intercept to $y_1 - y_4$ are all fixed at 1 to reflect the fact that the intercept has a constant effect on the dependent variable over time (this is the SEM analog to β_{0i} in the MLM equation having no multiplier). The fixed coefficients on the paths from slope to $y_1 - y_4$ take on the integer values 0 to 3 to reflect a linearly increasing effect of time (this is the SEM analog to β_{1i} in the MLM equation being multiplied by *time*). The triangle labeled “1” represents an intercept in the SEM framework. All other model parameters are directly analogous to MLM parameters, as shown in Figure 21.4A .

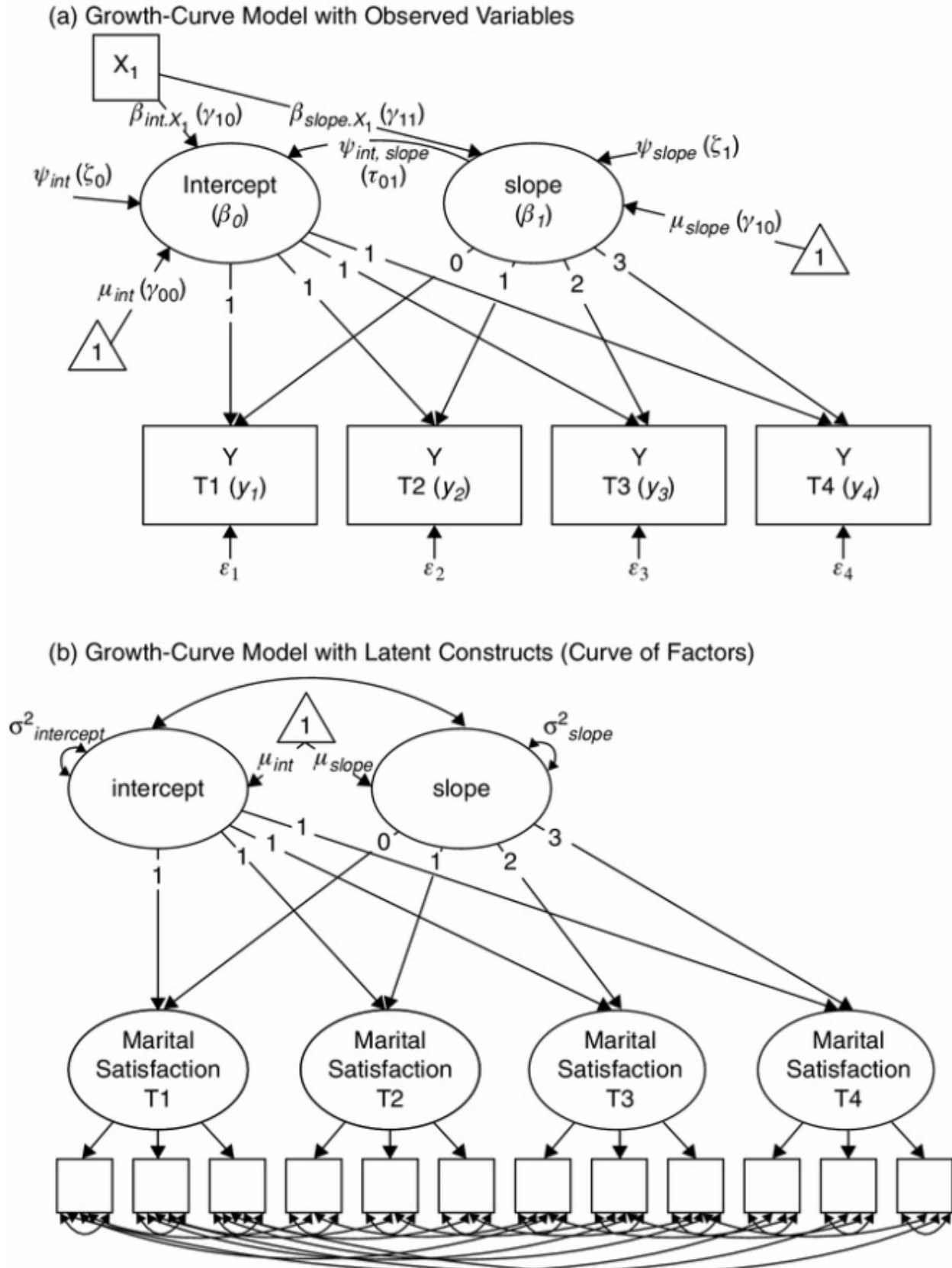


Figure 21.4. Growth curve models. The upper diagram (A) depicts a growth

curve model where the dependent variables are manifest (observed). The lower panel (B) depicts a model where the dependent variables are latent. The latent-variable model removes measurement error from the dependent constructs, allowing the pattern of change across time in the construct to be estimated with greater accuracy.

Interpreting latent growth curves. When interpreting the results from a latent growth curve, there are five main parameters of interest: the mean of the intercept, the mean of the slope, the variance of the intercept, the variance of the slope, and the covariance between the intercept and slope. The mean of the intercept (the fixed intercept in MLM) is the mean score at time 1 (or the time when the slope is equal to 0). The mean of the slope factor (or the fixed slope of time in MLM) is the mean change over time across all individuals in the sample. The variance of the intercept factor (or the variance of the random intercept in MLM) is the amount of variability in scores at time 1 across the sample, and the variance of the slope (or the variance of the random slope in MLM) is the variability in the rate of change over time across the sample. Finally, the covariance of the intercept and slope (or the covariance between the random intercept and random slope in MLM) represents the relation between individuals' starting points (intercepts) and their change over time. A positive covariance indicates that individuals who start out higher on the dependent variable have greater positive change (or less negative change) over time, whereas a negative covariance indicates that individuals who start out lower on the dependent variable have greater positive change over time. For example, data on participants' negative affect was collected over four time points, and we fit a linear growth curve to the data (for further details on this data, see Little, 2013). The model was estimated using both MLM and SEM. Parameter estimates are reported in [Figure 21.5](#) and they did not differ between MLM and SEM in the first two decimal places. The mean of the intercept was 2.09, meaning that participants had an average score of 2.09 on the negative affect measure at time 1. The variance of the intercept was significantly different from zero (.30), indicating that there was variability in participants' initial scores. The mean of the slope of negative affect was -0.10 , meaning that on average negative affect scores declined by 0.10 at each timepoint. The variance of the slope was significantly different from zero (0.02), indicating that there was variability in participants' rate of change. Finally, the covariance between the intercept and slope was $-.79$, meaning that participants with higher initial scores had slopes that were more negative (a faster rate of change).

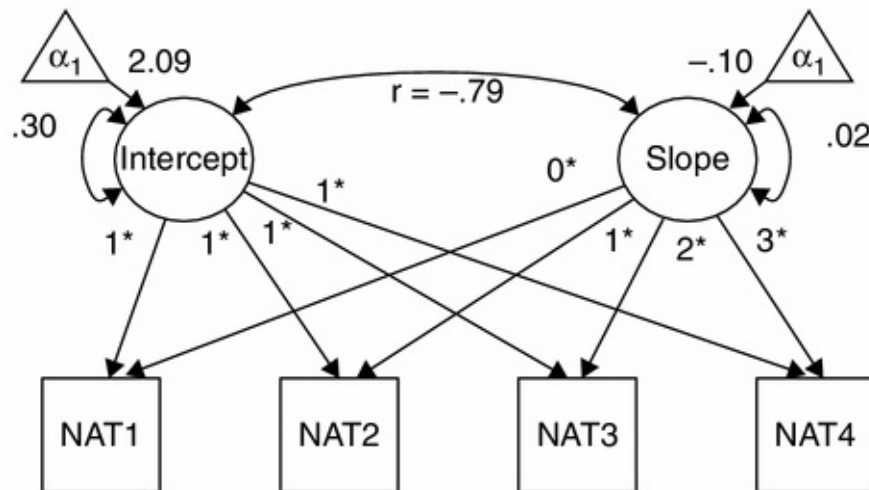


Figure 21.5. Example of a latent growth curve of negative affect estimated with both MLM and SEM. Parameters with a * are fixed in the model and not estimated.

Table 21.3. *Alternative Parameterizations of Growth Curves*

	Time 1	Time 2	Time 3	Time 4
Intercept at Time 1	0	1	2	3
Intercept at Time 4	-3	-2	-1	0
Intercept at Time 2	-1	0	1	2
Intercept between Time 2 and Time 3	-1.5	-.5	.5	1.5

Alternative parameterizations. When estimating latent growth curves, the location of the intercept is an important consideration. The mean of the intercept in a latent growth curve is interpreted as the mean of the dependent variable at the time point where the slope is set to 0. Table 21.3 contains four different ways to parameterize the slope variable. Each parameterization places the intercept at a different point in time, including placing the intercept between two measured time points. Depending on the hypotheses of interest, the placement of the intercept can be extremely important. For example, suppose students are given a skills test at four occasions while they are learning to play the guitar: the day they begin and then three more times at even one-month intervals. We might expect to see very little variance in the intercept if it is set as the first time point, as the novice students have not yet had an opportunity to distinguish themselves from each other; however, if the intercept is set as the fourth time point, there would be much greater variance in that intercept. Similarly, a researcher might

be interested in testing the rate of achievement growth up to the time of taking a standardized test. In this case, again, it would make most sense for the standardized test to be the reference point because its mean and variance are of more theoretical interest than occasions leading up to that time .

Comparing SEM and MLM. The key differences between assessing latent growth curves in SEM and MLM are the flexibility to add complexity to the overall model that is fit to the data. In both the SEM and MLM approaches the intercept and slope of the latent growth curve can be predicted by other variables. However, in the MLM approach, latent growth curve models are limited to a single growth curve, and the intercept and slope of a latent growth curve cannot predict other variables. In the SEM approach, the intercept and slope of latent variables can predict and be predicted by other variables and multiple growth curves can be simultaneously estimated. In SEM, it is also possible to replace the manifest outcomes with latent variables, as seen in [Figure 21.4B](#); this approach is sometimes called a curve-of-factors model, as the variables subjected to the growth curve are now latent factors rather than measured variables. Note that being able to interpret the curve-of-factors model requires establishing strong invariance. Another advantage of SEM is the possibility of allowing the model to estimate the coefficients of the slope factor rather than specifying these a priori. For example, rather than specifying coefficients of 0, 1, 2, 3 (or any of the options shown in [Table 21.3](#)), the modeler may specify that the first loading is 0 (to set the intercept), the last loading is 1 (to set the scale of the slope factor), and allow the other two occasions' loadings to be freely estimated. This freely estimated curve is called a "latent basis" curve (Singer & Willett, [2003](#)). In this way it is not necessary to constrain the growth trajectory to be linear or quadratic or to take any other prespecified form. For further information on latent growth curves, see Preacher *et al.* ([2008](#)), Singer & Willett ([2003](#)), or Little (2013).

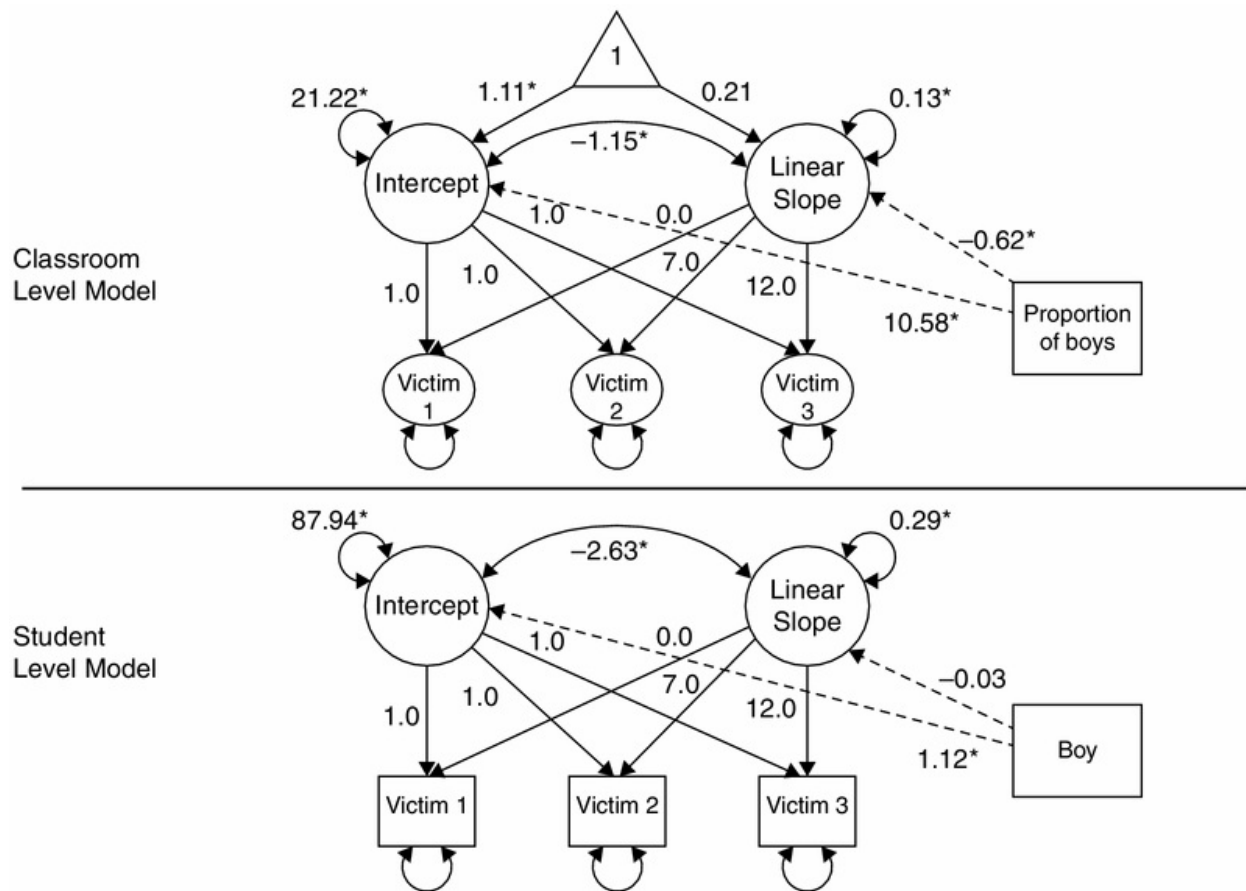


Figure 21.6. MSEM growth curve model. Growth curves are estimates at both the student and classroom levels. Intercept and slope means can only be estimated at the between (classroom) level. However, intercept and slope variance can be predicted at both levels. The growth curves are parameterized to represent change per month, thus the slope loadings represent the number of months between measurements.

MLMs and SEMs can be handled using a variety of software programs. Currently, programs fall into one of four categories: programs specifically focused on MLMs (e.g., HLM, MLwiN), programs specifically focused on SEM (e.g., AMOS, EQS), programs that can estimate both MLMs and SEMs (e.g., *Mplus*, LISREL), and general statistical programs that can estimate MLMs (e.g., SPSS, SAS PROC MIXED, R [nlme and lme4 packages], and STATA) or SEMs (SAS PROC TCALIS, R [lavaan, sem and OpenMX packages], and STATA). A complete review of SEM and MLM software is beyond the scope of this chapter; however, we wish to highlight the capabilities of the free, open-source software package R (R Development Core Team 2011). In particular, the lme4 (Bates, Maechler, & Bolker, 2011), openMX (Boker et al., 2011), and lavaan (Rosseel,

2012) packages provide powerful, intuitive options to analyze models in MLM and SEM. For a complete review of MLM software we recommend Snijders and Bosker (2011), and for a complete review of SEM software we recommend Kline (2010).

Multilevel SEM

When analyzing longitudinal data, both SEM and MLM have limitations (e.g., MLM can only have one outcome variable and that variable must be measured at level 1; SEM can only accommodate one level of nesting). Multilevel SEM (MSEM) is a new, flexible technique that has fewer restrictions than SEM or MLM. In MSEM, a structural equation model is estimated at both level 1 and level 2 (Lüdtke et al., 2008). This allows researchers to estimate a three-level model when one level of nesting is attributable to a longitudinal design. For example, if children are nested within classrooms over time, there is a three-level nested structure, with observations nested within children nested within classroom. This design could be analyzed by estimating a panel model or latent growth curve at both the child and classroom levels. Thus, MSEM allows researchers to estimate separate effects for children and classrooms.

For example, in a study assessing the KiVa program, an anti-bullying program developed in Finland, Williford, Boulton, Noland, Little, Kärnä, & Salmivalli (2011) measured 4,131 students nested within 211 classrooms (in the intervention condition) over 3 time points. Using MSEM, a latent growth curve assessed the change in bullying victims at both the individual and classroom levels with gender as a predictor of both the intercept and slope at both levels (Figure 21.6). In MSEM, means can only be estimated at the between level, thus in this model the intercept and slope only have means at the classroom level. However, the variance in intercepts and slopes can be divided into between (classroom) and within (student) levels. Furthermore, the intercept and slope can be predicted at both levels. In this model, gender (at the student level) and the proportion of boys in a classroom (at the classroom level) predicted the intercepts, with boys bullying more than girls and classrooms with higher proportions of boys having higher frequency of bullying. Gender did not predict change in bullying for students, but the proportion of boys in a classroom was negatively related to the slope at the classroom level. The higher the proportion of boys in a classroom, the greater the decline in bullying over time.

MSEM provides advantages over traditional multilevel models, including

using latent variables to represent constructs, allowing between-group variables as outcomes, modeling relations among multiple constructs at the same time, and obtaining unbiased estimates of effects within (e.g., student level) and between (e.g., classroom level) groups (Lüdtke et al., 2008). Furthermore, MSEM provides unbiased estimates of mediation effects in nested data (Preacher, Zyphur, & Zhang, 2010) .

Conclusion

In this chapter we have provided a broad overview of best practices for handling nested data structures. As we have emphasized here, nested data structures are no longer a statistical issue to be avoided but instead a venue to examine important theoretical relations at various levels of influence. With the two modern approaches to analyzing dependent data structures, researchers now have the tools to embrace the nuances of hierarchically nested data and to model the intrinsically interdependent nature of social life in the real world. In particular, as social-personality psychologists become increasingly interested in how variables and relationships change over time (using both short-and long-time scales), longitudinal modeling will be an important tool for social and personality psychologists. By alerting researchers to the various advantages and opportunities that these tools afford, we feel confident that innovative questions and important discoveries are now within easy reach of social-personality psychology researchers.

As is often the case, advances in methodology can stimulate richer and more nuanced research questions. The techniques that we have reviewed here are extremely flexible and adaptable to allow researchers to formulate and test complex models. We encourage researchers, new and established, to master these techniques and to use them to enlighten the field. By wedding complex models with the power of today's multilevel and longitudinal analyses techniques, researchers can readily address the layers of interdependence and influence that characterize human behavior and its development.

References

- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.

- Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375–42.
- Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., *et al.* (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, 76, 306–317.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.
- Cheung, G. W., & Rensvold, R. B., (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255.
- Cole, D. A., & Maxwell, S. E. (2003). Testing meditational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, 112, 558–577.
- Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, 47, 1231–1236.
- Grant, A. M., & Gino, F. (2010). A little thanks goes a long way: Explaining why gratitude expressions motivate prosocial behavior. *Journal of Personality and Social Psychology*, 98, 946–955.
- Hayes, A. F., Preacher, K. J., & Myers, T. A. (2011). Mediation and the estimation of indirect effects in political communication research. In E. P. Bucy & R. L. Holbert (Eds.), *The sourcebook for political communication research: Methods, measures, and analytical techniques* (pp. 434–465). New York: Routledge.
- Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.
- Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.
- Laurenceau, J., Barrett, L. F., & Rovine, M. J. (2005). The interpersonal process model of intimacy in marriage: A daily-diary and multilevel modeling approach. *Journal of Family Psychology*, 19, 314–323.
- Little, T. D. (2013). *Longitudinal structural equation modeling*. New York: Guilford Press.

- Little, T. D., Preacher, K. J., Selig, J. P., & Card, N. A. (2007). New developments in SEM panel analyses of longitudinal data. *International Journal of Behavioral Development*, 31, 357–365.
- Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13, 203–229.
- Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, 44, 764–802.
- Maxwell, S. E., Cole, M. A., & Mitchell, M. A. (2011). Bias in cross-sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behavioral Research*, 46, 816–841.
- McNulty, J. K., & Russell, M. (2010). When “negative” behaviors are positive: A contextual analysis of the long-term effects of problem-solving behaviors on changes in relationship satisfaction. *Journal of Personality and Social Psychology*, 98, 587–604.
- Pornprasertmanit, S., & Little, T. D. (2012). Determining directional dependency in causal associations. *International Journal of Behavioral Development*, 36, 313–322.
- Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: Sage Publications.
- Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15, 209–233.
- Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, 59, 413–425.
- R Development Core Team (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Newbury Park, CA: Sage Publications.
- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351–357.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1–36.
- Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social-Personality Psychology Compass*, 5/6, 359–371.
- Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, 7, 422–445.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Methods for studying change and event occurrence*. New York: Oxford University Press.
- Snijders, T., & Bosker, R. (2011). *Multilevel analysis* (2nd ed.) London: Sage Publications.
- Watson, C. B., Chemers, M. M., & Preiser, N. (2001). Collective efficacy: A multilevel analysis. *Personality and Social Psychology Bulletin*, 27, 1057–1068.
- Wendorf, C. A. (2004). Primer on multiple regression coding: Common forms and the additional case of repeated contrasts. *Understanding Statistics*, 3, 47–57.
- Williford, A., Boulton, A., Noland, B., Little, T. D., Kärnä, A., & Salmivalli, C. (2010). Effects of the KiVa anti-bullying program on adolescents' depression, anxiety, and perception of peers. *Journal of Abnormal Child Psychology*, 40, 289–300.

¹ If a model contains a random intercept or slope term with no corresponding fixed effect, then the fixed effect is assumed to be exactly 0 – an unrealistic assumption. However, a fixed effect with no corresponding random effect can

frequently be justified.

² In some cases a model with all residual covariances freely estimated may not be identified – for example, if the number of indicators per latent construct is only 2. In this case, residual covariances should not be fixed to 0, but they may be constrained to be equal at each time interval. That is, residual covariances between time 1 and time 2, time 2 and time 3, and time 3 and time 4 (lag-1) may be fixed to equal each other, and those between time 1 and time 3 and time 2 and time 4 (lag-2) may also be fixed to equal each other.

Chapter twenty-two The Design and Analysis of Data from Dyads and Groups

David A. Kenny and Deborah A. Kashy*

Much of social psychological theory and research studies people as individuals. This individualistic orientation is attributable in part to the fact that standard statistical methods, such as ANOVA and multiple regression, require independence between observations. Nonetheless, many social psychological concepts (e.g., attraction, imitation, person perception, helping, aggression, communication, and influence) intrinsically involve two persons, and several others (e.g., leadership, cohesiveness, productivity, conformity, norms, and group polarization) require groups of persons. Before we can have a genuinely *social* psychology, our theories, research methods, and data analyses must take into account the truly interpersonal nature of the phenomena under study.

Because the phenomena that social psychologists study are social by definition, observations from social data do not refer to a person, but rather to multiple persons embedded within a social context (Bond & Kenny, 2002). Consider, for instance, how much Harry likes Sally. Because the checkmark on a piece of paper is made by Harry, researchers all too often make the fundamental attribution error and treat the measurement as if it referred to only Harry. Almost certainly the liking that Harry feels for Sally is driven in part by characteristics of Sally herself, such as how friendly or agreeable she is, as well as by the unique relationship that Harry and Sally have established. The measurement reflects both Harry and Sally, and so it is fundamentally dyadic.

The intrinsically social nature of the measurements in a social psychology study implies that they are often linked to other measurements in the study, and the strength of these links may be one of the most important research questions to be examined. Consider the following examples:

The degree to which a perceiver sees two other persons as both being communally responsive was measured in 31 triads (Lemay & Clark, 2008, Study 5).

Members of 60 4-person groups interact and rate each other's leadership and

the extent to which members agree about who is the leader is measured (Dabbs & Ruback, 1987).

Romantic partners are observed interacting and each person's disclosure is assessed to examine reciprocity in 62 dating couples (Webster, Brunell, & Pilkington, 2009).

The degree of similarity in aggression exhibited by the members of 142 dyads is examined (Anderson, Buckley, & Carnagey, 2008).

The degree to which a family member is positively engaged with the other family members is measured in 445 4-person families (Ackerman, Kashy, Donnellan, & Conger, 2011).

In each of these cases, an important social psychological phenomenon – assimilation, agreement, reciprocity, similarity, and consistency – is the focus of study. Yet none of them can be easily addressed by standard methods of ANOVA and multiple regression. This chapter provides an introduction to methods that are appropriate for analyzing data in which observations are linked as a result of being in the same group or dyad.

In this chapter we begin with terminology and definitions that are essential for understanding dyadic and group data. We then discuss a statistical technique that has become indispensable for the analysis of such data – multilevel modeling (MLM). Next we provide a detailed discussion of dyadic data structures in which both members are measured on the same set of variables. For dyadic data, the discussion centers on the very popular Actor-Partner Interdependence Model (APIM). We then turn our attention to group data. We discuss first the extension of the APIM to the study of groups. We then discuss the analysis of data from groups in which each person rates or interacts with other members of the group, and finally we discuss the analysis of intergroup data. The final data structure, one likely unfamiliar to most readers but important, represents a blend of dyadic and group structures, the one-with-many design. In each section we emphasize that a person's response depends not only on that person but also on the partners with whom he or she interacts. We avoid presenting extensive computational and computer syntax details, but we do cite the sources that present these details.

Terminology and Definitions

There are several key terms that are frequently used in the discussion of the design and analysis of dyad and group data structures. They are nonindependence of observations, types of variables, and distinguishability. We

shall use these concepts extensively throughout the remainder of the chapter.

Independence of Observations

The key assumption of ANOVA and multiple regression is that once variation resulting from the predictor variables is controlled, the scores of different units are independent – that is, uncorrelated. However, social interaction almost by definition implies interdependence, which results in nonindependence. Social psychologists are keenly aware of the assumption of independence, and to ensure that it is not violated, they often have taken the approach of randomly assigning individuals to conditions and then testing them in isolation. If social interaction is necessary, they train confederates who behave in a predetermined or scripted fashion. In social cognition research, stimulus persons are often presented via computer. Indeed, even research that is explicitly intended to study dyadic and group processes often involves having research participants unwittingly interact with a preprogrammed computer as their “partner.” In fact, in a recent survey of papers published in the *Journal of Personality and Social Psychology* (Kashy & Donnellan, 2012), only 11% involved the study of actual dyads or groups. By eliminating nonindependence, social psychologists are to some extent taking the “social” out of their research.

Investigation of dyadic or group processes requires that social psychologists use research designs and analysis strategies that recognize the interdependence of social behavior. Social psychologists should treat interdependence not as a statistical nuisance that should be controlled, but rather as an important social psychological phenomenon that should be studied. Reciprocity, agreement, accuracy, and consistency all imply some form of nonindependence of data. Thus, our discussion of nonindependence focuses primarily on ways by which researchers can model interdependence between the thoughts, feelings, and behaviors of individuals imbedded within dyads and groups. All of the models which we discuss require statistical methods that allow for nonindependence (i.e., methods for which violations of the independence assumption does not bias statistical tests), and they also provide estimates of the degree of nonindependence attributable to dyads or groups.

Types of Variables

In dyadic and group research, variables can vary within dyads or groups, between dyads or groups, or both within and between dyads or groups (i.e., a “mixed” variable). A within-groups variable varies across the group members,

but when averaged across the individuals within the same group, each group has the same average score. Gender would be a categorical within-groups variable in a study of four-person groups in which each group is comprised of two men and two women. Indeed, gender is the prototypical within-dyads variable in studies of heterosexual marital relationships, because every couple is comprised of both a man and a woman. An example of a continuous within-groups variable might be percent of the time talking in four-person groups, because individuals within the group would vary in how much they participate in the conversation, but the average percent of the time each member talked in each group would always be 25%.

In contrast, a between-groups variable varies from group to group (or dyad to dyad) but does not vary across individuals within the same group. In other words, all members of the group or dyad have the same score on this type of variable. Task ambiguity (high, moderate, low) might be an example of a categorical between-groups variable in a study of group productivity if each group is randomly assigned to one of these three conditions. In dyadic research, examples of between-dyads variables would include number of years married or gender in research on gay and lesbian couples.

When variables vary both within and between dyads or groups, they are called *mixed* variables. Many variables in dyadic research are mixed variables in that the dyad members' scores may differ for the two partners and some dyads have higher average scores than other dyads. Likewise, in group research a mixed variable varies from person to person within the group and there is also variation from group to group in the average values. Motivation would likely be a mixed variable in research on group functioning because group members may differ from each other in motivation and average motivation may be higher in some groups than in others. Relationship trust is an example of a mixed variable in research on romantic partners because the two partners may differ from each other in trust levels, and some couples would have higher average levels of trust than other couples. Gender can also be a mixed variable if some dyads or groups are same-sex and others are mixed-gender (e.g., a study of gay, lesbian, and heterosexual couples; Goldberg, Smith, & Kashy, 2010). All the outcome variables that we discuss in this chapter are mixed variables.

It is also helpful to consider the level of measurement of the outcome variable. Measurements in dyadic research can be obtained at the level of the individual such that each of the two members has his or her own score. For example, in a study of dating couples, the number of affectionate comments directed toward

the partner would be an individual-level measure that would be classified as a mixed variable. Measurements can also be obtained at the level of the dyad such that each dyad has only one score, and so the outcome variable is between-dyads. For example, the physical distance between two persons would be a dyad-level measure. When an outcome is measured at the dyad level, then dyad should be treated as the unit of analysis that basically involves analyzing the data as if dyad were the “participant.” In contrast, when outcomes are measured for each individual, the statistical approach needs to model the nonindependence between dyad members’ scores, the key topic of this chapter.

In a parallel fashion, data from group research can be measured at the level of the group, as would be the case if the outcome measure were a single index of group productivity. As with dyadic data, if only a single outcome is obtained for each group, then group is treated as the unit of analysis, and special analytic methods are not required. However, data from group research can also be measured at the individual level so that each person within the group contributes a single score. An example of an individual-level measure for group data is when each group member rates his or her satisfaction with the group.

Group research also allows for a third type of measurement. Specifically, measurements can be obtained at the dyad level within a group so that each individual is paired with each of the other group members. For example, in a 4-person group (persons A, B, C, and D) where each individual rates every other individual in the group on a measure of friendliness, there would be 12 dyadic ratings (person A rating person B or AB, AC, AD, BA, and so on). This type of dyadic-level group data results in a fairly complex web of nonindependence but can address several important social psychological questions . (See the Social Relations Model section later in this chapter.)

Distinguishability

Another important consideration is whether all the dyad or group members can be systematically distinguished from one another as a function of some variable. Members of a dyad or group are conceptually distinguishable if there is a meaningful variable that can be used to differentiate all the individuals with each of the dyads or groups. For example, researchers who study heterosexual marital relationships typically treat gender as a distinguishing variable: husbands versus wives. In studies of families, individuals are typically distinguished by their family role – mother, father, older sibling, younger sibling. Distinguishability is an important consideration in the analysis of data from dyads and groups

because it suggests that there may be systematic differences in both processes and outcomes for individuals in the different roles.

In contrast, dyad or group members are often indistinguishable or exchangeable in that there is no systematic way to order the scores. Examples of indistinguishable dyads include homosexual couples or identical twins, and examples of indistinguishable groups include groups formed in the laboratory by randomly assigning research participants to groups. Indistinguishability implies that the individuals within the dyad or group are sampled from the same underlying population. As a result, data-analytic models for distinguishable data often differ from those for indistinguishable data. Kenny, Kashy, and Cook (2006) present methods for testing whether dyad members who are conceptually distinguishable are also empirically distinguishable using structural equation modeling (SEM). In the dyad data structures section of this chapter, we briefly discuss how to do so using MLM.

When we have measurements on the same variable for each dyad or group member, standard methods usually fail to capture the nonindependence in the data. To handle that nonindependence, we can use MLM, the topic to which we now turn .

Multilevel Modeling

The past 10 to 15 years have seen the development and popularization of MLM. It represents a generalization of ANOVA and regression methods that allows for the nonindependence inherent in dyadic and group research. It also can estimate the effects of within, between, and mixed independent variables. Therefore, we have chosen to focus this chapter on how MLM can be used in dyadic and groups research. SEM is almost always an alternative estimation approach for the designs that we discuss. In fact, when members are distinguishable, it may be easier to use SEM than it is to use MLM. Moreover, some models can be estimated only by SEM. However, in this chapter, we discuss only MLM because it is easier to implement, more similar to ANOVA and multiple regression, and MLM software is typically more accessible than SEM.

MLM (also known as hierarchical linear modeling, mixed models, and random coefficient estimation) is an important tool for researchers in social psychology and is especially useful for researchers who study interpersonal behavior in dyads and groups. (It is also especially important for researchers who study change; see Schoemann, Rhemtulla, & Little, Chapter 21 in this

volume, for this type of MLM application.) A complete treatment of MLM is beyond the scope of this chapter, and so we refer readers, in addition to the aforementioned Schoemann *et al.* chapter, to several texts such as Hox (2010), Raudenbush and Bryk (2002), and Snijders and Bosker (1999). Also, Kenny *et al.* (2006) and Kenny and Kashy (2011) provide an expanded treatment for dyadic analyses with particular reference to MLM. MLM software is available in stand-alone programs (e.g., HLM and MLwiN) or embedded in commonly used computer packages (e.g., SPSS, SAS, Stata, or R).

In MLM terminology, groups or dyads are considered to be the *upper-level* or *level two* units, and individuals who are nested within the groups are treated as *lower-level* or *level one* units. To be appropriate for analyses using MLM, outcome scores must be measured for each lower-level unit, and predictors can be measured at either the upper or lower levels. In our brief introduction, we first describe a typical multilevel analysis for group data, and then we note how this analysis must be modified to accommodate dyadic data. Our initial discussion of MLM presumes that group or dyad members are indistinguishable, and in a subsequent section we present the alternative specification for the distinguishable case for dyads.

MLM and Group Data

As an example, consider a study by Weingart, Brett, Olekalns, and Smith (2007) that investigated the effects of cooperative versus individualistic motivations on flexibility in negotiation strategies. The study included 144 management students who were assigned to 36 4-person groups. In this example, group is the upper-level unit and individual is the lower-level unit. The students were from one of two universities, and so we have a variable that we call University. Because all members of each group were from the same university, this variable is a between-groups variable; in MLM terms, University is an upper-level predictor variable. In this study, each individual within a group had either a cooperative or individualistic orientation, a variable we call Orientation. Orientation is a mixed predictor variable as it varies from individual to individual within most groups, and some groups have more persons with a cooperative orientation than others do. In MLM terms within-groups variables and mixed variables such as Orientation are lower-level predictors. The outcome, Flexibility, is collected for each individual or each lower-level unit.

Note that in this example both predictors are categorical and each needs to be coded using either dummy coding or effects coding. We use effects coding such

that university A is given a negative one (-1) and university B is given a positive one (1); likewise a cooperative orientation is given a $+1$ and an individualistic orientation is given a -1 . In MLM, to be able to interpret the results zero needs to be a meaningful value for all predictor variables, both lower and upper level. For the example, we used effects coding, making zero a point halfway between the two categories. We could have used dummy coding, and the interpretation of the intercept would have to change accordingly.

MLM analyses typically require that the data are structured such that there is a separate record for each lower-level unit (i.e., person). Thus, each four-person group generates four data records, and the total number of rows or records for the entire study equals the total number of participants, which would be 144 in the example. Table 22.1 presents an illustration of the data that might be generated by this study. A critical component of the data set is that each individual's data must include a variable that denotes group membership so that members of the same group can be linked together for the analysis. The file also contains a person identifier as well as the variable "Member," which designates the four persons by 1 through 4. If group members are indistinguishable, assignment of an individual to particular Member number (e.g., 3) is totally arbitrary. The Member variable is not required but can be helpful in some analyses. Each data record also should include the person's outcome score (e.g., Flexibility), any lower-level predictor variables (e.g., Orientation), and any upper-level predictor variables (e.g., University). Note that in the table the values of the upper-level predictor variable are repeated for each individual within the same group, illustrating that upper-level variables are what we have described as between-groups variables.

Table 22.1. Hypothetical Data for the Weingart et al. (2007) Study

Person	Group	Member	University (X)	Orientation (Z)	Flexibility (Y)
1	1	1	1	1	8
2	1	2	1	1	5
3	1	3	1	-1	6
4	1	4	1	-1	6
5	2	1	1	1	5
6	2	2	1	1	5
7	2	3	1	1	8
8	2	4	1	-1	5
9	3	1	-1	1	9
10	3	2	-1	-1	7
11	3	3	-1	-1	6
12	3	4	-1	-1	5

MLM is often described as a multistep process in which a lower-level regression is first computed separately for each upper-level unit. In the current example, this would imply that a separate regression equation is computed for each group, and in each of these regressions the individuals' flexibility (Y) is predicted as a function of that individual's orientation (X). More formally, the lower-level model for person i in group j would be

$$Y_{ij} = b_{0j} + b_{1j}X_{ij} + e_{ij}$$

where b_{0j} represents the mean flexibility score for group j averaged over the levels of orientation, and b_{1j} represents one half of the mean difference between the two orientation types for group j . The next step of the multilevel analysis involves treating the slopes and intercepts from the first-step analyses as "outcome" variables in two different level-two regressions. For these level-two analyses, the regression coefficients from the first step are assumed to be a function of our group-level predictor, University or Z . These two equations would be:

$$b_{0j} = a_0 + a_1 Z_j + d_j$$

$$b_{1j} = c_0 + c_1 Z_j + f_j$$

The first equation predicts the intercepts from the lower-level regressions as a function of the upper-level variable, Z . Because we have two effects-coded variables, in the example, a_0 is the grand mean of flexibility averaged over the two universities and the two levels of orientation, and $2a_1$ estimates how much more or less flexible students from university A were relative to students from university B. We multiply a_1 by 2 because that represents the difference in the Z s for persons at the two universities (+1 to -1). The second equation predicts the slopes from the lower-level regressions as a function of Z . Here c_0 estimates whether individuals with cooperative versus individualistic orientations differ in flexibility, and c_1 estimates the degree to which the effect of orientation on flexibility differs by university. These four coefficients, a_0 , a_1 , c_0 , and c_1 , are called *fixed effects* because they reflect average relationships among the study variables.

There are three random effects that are the error terms represented in these three equations. First, there is the variance in the error component, e_{ij} , in the

level one equation or σ_e^2 , which represents variation in responses across individuals, after controlling for the effects of the lower-level predictor variable(s). In the example, this component represents variation in flexibility from individual to individual within a group, controlling for the effects of orientation. There are also random effects in each of the two upper-level regression equations. The random effect d_j in the first of these regression equations represents variation in the intercepts that is not explained by Z . For the example, d_j represents variation in the group means for flexibility that is not explained by university, and so this component accounts for the nonindependence due to groups. The variance in the estimated d_j 's is a combination of σ_d^2 , which can be referred to as the *group variance*, and σ_e^2 . The intraclass correlation which measures the relative degree of nonindependence in flexibility due to groups can then be defined as

$$\rho = \frac{\sigma_d^2}{\sigma_d^2 + \sigma_e^2}$$

The random effect in the second of the upper-level regression equations is f_j and represents the degree to which the size of the orientation-flexibility relationship varies from group to group after controlling for university. The variance in f_j is a combination of σ_f^2 and σ_e^2 where σ_f^2 can be referred to as the variance in the orientation-flexibility slopes. In many applications of MLM to small-group contexts, this variance can be very small and not statistically different from zero. (If this is the case, the random slope would be removed from the model.) Finally, the random effects for the intercept and slope may covary (e.g., if a group is relatively high in flexibility, is the link between orientation and flexibility stronger for that group?), adding a fourth parameter of random effects to the model.

Conceptualizing MLM as two-step procedure is useful when introducing MLM. In fact, most MLM programs do not use least squares estimation but rather a form of maximum likelihood estimation. The procedure is iterative, and some models (e.g., if σ_f^2 is small) can be difficult to estimate.¹ Most MLM programs estimate the model parameters using a single combined equation that can be derived by plugging the two upper-level equations for b_{0j} and b_{1j} into the lower-level equation:

$$Y_{ij} = (a_0 + a_1 Z_j + d_j) + (c_0 + c_1 Z_j + f_j) X_{ij} + e_{ij}$$

which can be rearranged as:

$$Y_{ij} = (a_0 + d_j) + a_1 Z_j + (c_0 + f_j) X_{ij} + c_1 Z_j X_{ij} + e_{ij}$$

The intercept in this combined equation involves two components: a_0 is the fixed effects piece that estimates the grand mean of flexibility, and d_j is the random effects piece that indicates that average flexibility also varies from group to group. As we have noted, it is the variation in d_j that measures group nonindependence. The main effect of Z is comprised of only a fixed effect represented by a_1 . Given the coding scheme, it is half of the mean difference in flexibility between the two universities. The main effect of X has both a fixed effect component, c_0 , which measures the average orientation-flexibility relationship across all of the groups, and a random effect component, f_j which estimates the degree to which the effects of orientation on flexibility varies from group to group. The coefficient for the interaction of X and Z , c_1 , is the last fixed effect and it estimates the degree to which the orientation-flexibility relationship is moderated by university. The final term in the model is the random error component for flexibility ratings: e_{ij} .

MLM and Dyadic Data

In research with dyads, the “group” size is limited to two. The small number of lower-level units (individuals) nested within the upper-level units (dyads) necessitates one important simplification of the multilevel model: The random component for the slopes associated with the lower-level predictor, f_j must be fixed to zero. In other words, the three separate MLM equations are

$$Y_{ij} = b_{0j} + d_{1j} X_{ij} + e_{ij}$$

$$b_{0j} = a_0 + a_1 Z_j + d_j$$

$$b_{1j} = c_0 + c_1 Z_j$$

and the single combined equation for dyadic data is:

$$Y_{ij} = (a_0 + d_j) + a_1 Z_j + c_0 X_{ij} + c_1 Z_j X_{ij} + e_{ij}.$$

Thus, the only difference in the case of dyads rather than larger groups is that the slopes for the level one variable do not vary – in other words, there is no error

term in the b_1 equation. (The model could not be estimated if we allowed slopes to randomly vary.) However, the random intercept or d_j remains, and as with group data, the presence of this random effect in the intercepts models the nonindependence of scores for the two dyad members. Similarly, the intraclass correlation can be estimated as the proportion of variance in these intercepts or $\sigma_d^2/(\sigma_d^2 + \sigma_e^2)$.

Negative Nonindependence

In the standard MLM formulation, nonindependence is modeled as a variance in the intercepts. That is, to the extent that there is more variance in the average responses across the upper-level units or groups, relative to the variance from person to person within the groups, there is evidence that the scores within the groups are correlated. This method of specifying nonindependence is generally adequate for most outcomes from group research. However, it is possible for nonindependence to be negative such that outcome scores within the same group are more different from one another than are scores from two different groups. This situation can occur when the outcome is structured such that when one group member has a high score (e.g., talks a great deal during a group interaction), by necessity other group members have low scores (e.g., they do not get much talk time). Indeed, theoretically the intraclass correlation can actually range from +1.0 to $-1/(n - 1)$ where n is the group size. So in 4-person groups the possible range of the intraclass correlation is from 1.00 to $-.33$, and in dyads the possible range is 1.00 to -1.00 .

If the intraclass correlation is negative but the nonindependence is modeled as a variance, data analysis programs have difficulties, often generating an error message. An alternative and more general specification is to not have a random intercept component in the model (i.e., the earlier mentioned s_d^2 term is dropped), but rather to have the errors of members in the same group be correlated. This correlation of errors models the nonindependence, and because that correlation can be negative, negative nonindependence is possible. To allow for the correlation between errors, the MLM program needs to have some sort of repeated measures option. This formulation is also useful when group members are distinguishable because it then becomes possible to allow the error variances to differ by role. In sum, the standard MLM formulation can be problematic, particularly in studies of dyads, and we encourage researchers to consider using the correlated errors formulation rather than the random intercept approach when

analyzing dyadic data. If the nonindependence is positive, then this covariance in the residuals equals the variance of the intercepts or σ_d^2 , which was described earlier.

Dyadic Data and the Actor-Partner Interdependence Model

Dyadic data implies a study in which pairs of people are measured on the same set of variables, what Kenny *et al.* (2006) call the *standard dyadic design*. By far the most popular model for such data is the Actor-Partner Interdependence Model (APIM) and it is that model that we detail here. There are alternative models for dyadic data, most notably the Common Fate Model (Ledermann & Kenny, 2012), but these models are infrequently applied by practitioners.

The APIM focuses on an important ramification of nonindependence in dyadic and group research: One person's score on a predictor or causal variable may influence not only that person's score on an outcome variable, but also that person's partner's score on the outcome variable. This model can be applied to both dyadic and group research, and it can be used in research both when members are distinguishable (e.g., married couples) and when members are interchangeable (e.g., same-sex friends). Here we discuss the dyadic context and the extension to group research follows in the next section.

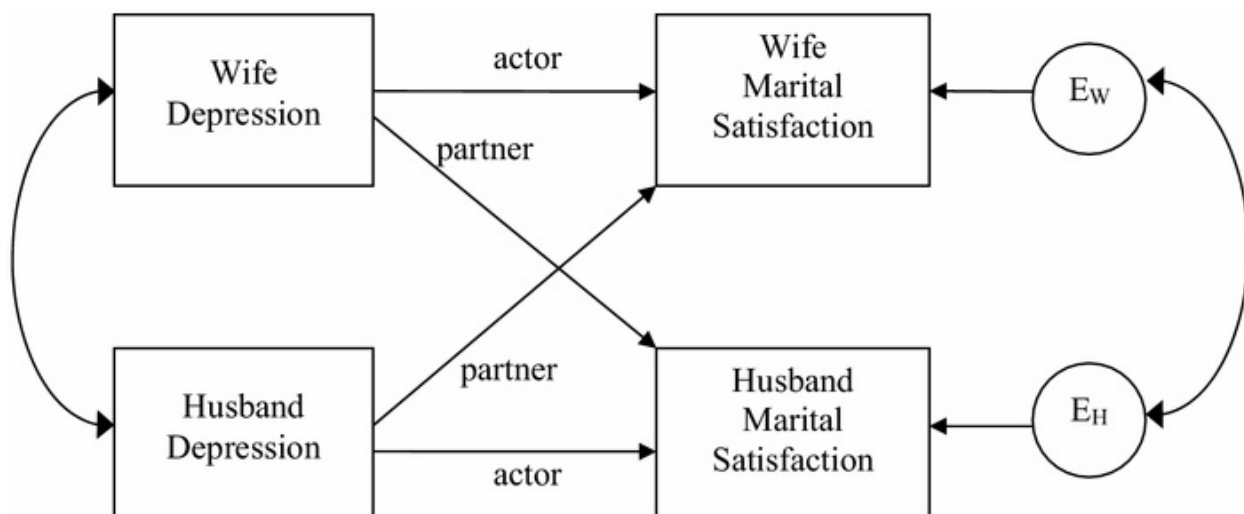


Figure 22.1. Example of actor and partner effects in the APIM.

As an example, consider the effects of depression on marital satisfaction for husbands and wives. As seen in Figure 22.1, it may be that a wife's depression

influences her own marital satisfaction as well as her husband's marital satisfaction. The effect of a wife's depression on her own marital satisfaction is called an *actor effect*, and the effect of her depression on her husband's satisfaction is called a *partner effect*. That is, an actor effect occurs when a person's score on a predictor variable affects that person's score on an outcome variable; a partner effect occurs when a person's score on a predictor variable affects his or her partner's score on an outcome variable. Note that to estimate both actor and partner effects, the predictor variable must be mixed.

The APIM can be estimated by either MLM or SEM (Kenny et al., 2006). In this chapter we focus on MLM. Consider first dyads whose members are indistinguishable. We have Y as the outcome, say relationship satisfaction, and X as the predictor, say depression. Both X and Y variables are mixed, and we can write two equations for the two members, 1 and 2, for dyad j :

$$\begin{aligned} Y_{1j} &= b + aX_{1j} + pX_{2j} + e_{1j} \\ Y_{2j} &= b + pX_{1j} + aX_{2j} + e_{2j} \end{aligned}$$

Nonindependence is modeled by the correlation between the e 's, and so the model includes three fixed effects, b , a , and p , as well as two random effects, the estimate of the error variance and the correlation of the errors. (Alternatively, we could have estimated the random intercept model, but especially for dyadic research we prefer the correlated errors approach.)

To estimate the APIM using MLM, we need what is called a pairwise data set (Kenny et al., 2006), which is a data set that has separate records for each dyad member and which also has both the person's and the partner's data on each record (see Table 22.2). Each record of the data set contains the dyad number (obviously no more than two records per dyad), a Member variable (1 or 2), the outcome variable score, Y_{ij} , the actor variable or X_{ij} , and the partner variable which we designate as X'_{ij} . Note that for a given dyad one person's actor variable is the other person's partner variable, as is shown by the boxes in Table 22.2.

Table 22.2. Data Depicting the Pairwise Data Structure with Indistinguishable Dyads

Dyad	Person	Member	Satisfaction (Y)	Actor Depression (X)	Partner Depression (X')
1	1	1	24	3	2
1	2	2	32	2	3
2	3	1	51	4	3
2	4	2	18	3	4
3	5	1	19	4	6
3	6	2	23	6	4
4	7	1	38	2	3
4	8	2	33	3	2
5	9	1	26	5	2
5	10	2	52	2	5

The single MLM equation for person i in dyad j is:

$$Y_{ij} = b + aX_{ij} + pX'_{ij} + e_{ij}$$

We allow for nonindependence by allowing for a correlation between the e_{ij} values across dyad members.

If dyad members are distinguishable, then there is a dichotomous within-dyads variable, say G_{ij} (e.g., to represent gender for example if the study were one of heterosexual married couples), which is an effects-coded (+1 and -1) variable. We add G_{ij} to the model and allow it to also interact with actor and partner:

$$Y_{ij} = b + aX_{ij} + pX'_{ij} + cG_{ij} + gG_{ij}X_{ij} + hG_{ij}X'_{ij} + e_{ij}$$

Given that we have used effects coding for G , b is the average intercept across the whole sample, and $2c$ estimates the mean difference for the distinguishing variable (e.g., the difference between husbands and wives). Likewise, a is the average actor effect and d estimates whether the size of the actor effect differs across the distinguishing variable. Finally, p is the average partner effect and g estimates the difference in partner effects. Thus, the actor effect when $G_{ij} = 1$ equals $a + g$ and when $G_{ij} = -1$ equals $a - g$, and the partner effect when $G_{ij} = 1$ equals $p + h$ and when $G_{ij} = -1$ equals $p - h$. Note too that the error variance may be heterogeneous in that it may take on different values for the two distinguishable members. Readers can consult methods to directly estimate the two actor and two partner effects using MLM in Kenny *et al.* (2006).

In sum, when we allow for distinguishability, we have four extra parameters, three of which are fixed effects – c , g , and h – and one random effect – a second residual variance. We can test whether the fit of model improves when we allow

for distinguishability. To do so, we estimate both the distinguishable and indistinguishable models using maximum likelihood estimation (not restricted maximum likelihood estimation, the default for most MLM programs), subtract the models' deviances (also called minus two times the model's log likelihood), and that difference under the null hypothesis that the dyads are indistinguishable is distributed as chi square with four degrees of freedom (Kashy & Kenny, 2012). Thus, even if a variable such as gender is theoretically a distinguishing variable, it is nonetheless important to establish that it is empirically necessary.

Finally, like any two variables, actor and partner variables may interact. An interaction term might be formed in the usual way by creating a product variable, for example, $X_{ij}X'_{ij}$. Alternatively, the interaction term might be formed as the absolute difference, for example, $|X_{ij} - X'_{ij}|$, in order to test hypotheses about similarity. Kenny *et al.* (2006) describe how these and other forms of actor-partner interactions can be estimated and interpreted.

Patterns

An important feature of an APIM study is the relative sizes of actor and partner effects. Following Kenny and Cook (1999), there are four major patterns of APIM effects. The first is actor-only in which the actor effect is non-zero and partner effect is zero. The second is partner-only in which the partner effect is non-zero and actor effect is zero. The third is the couple model in which the actor and partner effects are equal. With this pattern, the operative cause of Y_{ij} is the sum or average of the dyad members' predictors, $X_{ij} + X'_{ij}$. The fourth is the contrast model in which the actor and partner are equal in absolute magnitude but opposite in sign. With this pattern, the operative cause of Y_{ij} is the difference between the dyad members' predictors or $X_{ij} - X'_{ij}$.

Kenny and Ledermann (2010) show how these different patterns can be captured by the parameter $k = p/a$, or the partner effect divided by the actor effect. Note that when k is zero, we have the actor-only pattern, when k is 1, we have the couple pattern, and when k is -1 , we have the contrast pattern. If we have a partner-only pattern, Kenny and Ledermann recommend computing k as a/p . A bootstrap confidence interval for k can be useful. For instance, if that interval goes from 0.8 to 1.3, then the data are consistent with a couple pattern. As we discuss later, k can be helpful in tests of mediation and moderation. We note that the estimation of k and its confidence interval currently requires the use of SEM (Kenny & Ledermann, 2010).

Example

The APIM has been used in dyadic research hundreds of times. One example of an APIM study is Klumb, Hoppmann, and Staats (2006), who studied the effect of housework on stress levels measured using cortisol levels. They studied 52 German dual-earner couples with at least one child. The actor effect was positive and statistically significant: Doing more housework yourself leads to higher stress levels. Interestingly, the partner effect was negative and about the same size as the actor effect, indicating that the more housework was done by the partner, the lower the person's stress levels, although this effect was not statistically significant. The pattern of results suggest a contrast effect: One feels less stress if one does less housework than one's partner.

Mediation and Moderation

Mediation and moderation is relatively complicated in the APIM in that when dyad members are distinguishable and the mediator or moderator is mixed, there are four effects that each can be mediated or moderated by two mediators. The reader can consult Lederman, Macho, and Kenny (2011) for more information on mediation and Garcia, Kenny, and Lederman (2013) for more information on moderation. We briefly discuss one published illustration of mediation and one of moderation.

Studying heterosexual couples, Riggs, Cusimano, and Benson (2011) found that one's own attachment anxiety mediated the effect of one's own childhood emotional abuse on both one's own and one's partner's dyadic adjustment. The k value (i.e., partner divided by actor effect) from emotional abuse to attachment anxiety was near zero, consistent with an actor-only model, and the k for attachment anxiety to dyadic adjustment was near 1, indicating a couple-level model. Interestingly, this basic pattern emerged for both men and women.

In a moderation example, Cillessen, Jiang, West and Lazkowski (2005) examined same-gendered friends and estimated actor and partner effects for self-rated prosocial behavior on friendship security. The moderator is gender of the pair, a between-dyads moderator. When the friends were female, the actor effect was stronger than when the friends were male. The partner effect did not appear to be moderated by the gender of the friendship pairs .

Group Studies

Groups present special theoretical and methodological issues that do not occur for dyads (Moreland, 2010). Here we consider three topics in the analysis of group data: the extension of the APIM to groups, group studies with a dyadic outcome, and intergroup studies.

Extension of the APIM to Groups

Before describing the extension of the APIM to groups, we describe the traditional MLM analysis of group data (e.g., Raudenbush & Bryk, 2002). We return to the earlier-introduced study by Weingart *et al.* (2007), who studied the effect of orientation (cooperative vs. individualistic) on the flexibility of negotiation strategy. In this traditional model for group data, the lower-level predictor variable is averaged across *all* of the group members to create an upper-level predictor. In the example, this would be accomplished by averaging the orientation scores across all of the group members. The analysis would then model each person's flexibility as a function of his or her own orientation, at the lower level, as well as the average orientation for the group, at the upper level. A key question addressed in this analysis is the extent to which the effect of orientation occurs at the lower level (person – i.e., individuals who are more cooperative are more flexible) versus upper level (group – i.e., groups in which there are larger numbers of cooperative individuals tend to be more flexible on average). We could also allow for the two variables to interact. Finally, to allow for group nonindependence, the intercept is treated as a random variable.

The APIM extension of the dyadic model takes a different tack to the estimation of the “group” effect. A person's outcomes are considered to be a function of both one's own inputs and the average of one's partners' inputs, not the inputs of the entire group. Therefore, the actor effect in the group APIM has the same definition as in the APIM for dyads: It is the effect of the person's own inputs or attributes on his or her own outcomes. The partner effect has a somewhat different definition relative to the dyadic case. For groups, the partner effect is the effect of the average of the other group members' inputs or attributes on the person's outcome. Thus, for Weingart *et al.* (2007), the partner effect would be the effect of the average cooperation orientation of the other three members of the group on the person's flexibility. Note that unlike the group mean, the partner effect is a lower-level variable as the partner effect differs for each person. Group researchers typically examine actor effects but often ignore partner effects. Yet it is the partner effects that capture interdependence within groups because the presence of partner effects indicates that a person's response

depends on characteristics and behaviors of other group members. In the discussion that follows, we presume that group members are indistinguishable. The interested reader can consult Kenny, Mannetti, Pierro, Livi, and Kashy (2002) for a discussion of the analysis when dyad members are distinguishable.

The APIM has not been nearly as popular for the analysis of group data as it has been for the analysis of dyadic data. The reasons for this are not entirely clear, but they likely include the fact that the partner is a single individual in dyad research but not in group research, and that computations are a bit more complex for group data structures relative to dyad data structures. Nonetheless, the model has been used, and we consider two examples from the published literature.

Bonito (2002) studied 10 groups with 4 persons in each. He examined participation in groups in a task that involved typing. Being a good typist had a positive actor effect: Better typists participated more. However, a partner effect was also present but of the opposite sign: Having better typists in the group led the person to participate less. The pattern is therefore a contrast pattern: One participated more if one is a good typist and the others are not. In other words, participation is determined by one's skill relative to that of other members in the group.

Consider next a study of commitment in small personal growth groups (Kivlighan, Kivlighan, & Cole, 2012). The study is longitudinal, but we ignore that complication here. In this study, the researchers were interested in determining whether group absence norms predicted an individual's commitment to the group, operationalized as whether the individual attended a subsequent group session. Thus, the key predictor variable is attendance at a previous session, and the actor effect estimates whether being absent for a session predicts the individual's absence at the subsequent session. The actor effect therefore represents consistency in a person's behavior and is referred to as *commitment* by Kivlighan *et al.* (2012). The partner effect examines whether the average of the previous absences of the other group members predicts the individual's attendance. Kivlighan *et al.* (2012) found evidence of both actor and partner effects such that individuals who missed an earlier session were more likely to be absent, and individuals whose other group members missed an earlier session were also more likely to be absent. Interestingly, the magnitude of actor and partner effects was nearly equal, with the partner effect slightly larger than the actor effect.

As with dyadic data, in addition to estimating actor and partner effects, the

APIM for groups can be used to estimate the effects of an actor-partner interaction. Actor-partner interactions suggest a synergy between actor and partner effects, and in the Kivlighan *et al.* (2012) study, the researchers examined the multiplicative interaction between actor and partner effects. They found that the effects of group norms (i.e., partner effects) were stronger for more committed individuals. That is, the behavior of the other group members had a stronger effect when the individual tended to attend the meetings, but other group members' behavior had a weaker effect when the person did not attend meetings.

One option for the APIM with groups that is not possible with dyads is that the actor and partner effects can be allowed to vary by group. For instance, Kivlighan *et al.* (2012) could have allowed the effect of prior attendance of the actor and the partners to vary by group. By doing so, actor and partner become random variables and in essence their effects are allowed to interact with group. For Kivlighan *et al.* (2012), if the partner effect was random, we would know that the strength of norms varied by groups.

We have described the APIM generalization to groups as a straightforward generalization of the dyadic model. Recently, Kenny and Garcia (2012) have proposed additional interactions that can be tested only in groups but not in dyads. These interactions assess the extent to which (a) the actor is similar to others in the group and (b) other members in the group are similar to each other. These two interactions can be combined to create a diversity effect that captures the overall diversity of members. This model is called the Group Actor-Partner Interdependence Model or GAPIM. They also show how the model can be adapted to study between-groups (a single score for each group) and dyadic outcomes. In the next section, we examine dyadic outcomes in group research .

Dyadic Outcomes: The Social Relations Model

In some studies all, or a subset of all, the possible dyads from the group are created. Then data are gathered from both members of each dyad in the group. If data are available from all dyads in a group, then the design is a round-robin design, where each group member rates or judges for every other group member. For example, if the group includes four individuals (A, B, C, and D), a round-robin data collection would involve 12 dyadic measures (A rating B or AB, AC, AD, BA, BC, BD, and so on). Consider the following studies:

Lam, Van der Vegt, Walter, and Huang (2011, Study 2) asked the 4

members of 31 teams to make social comparison and harm judgments of fellow team members.

Finkel and Eastwick (2008) investigated 15 speed-dating events involving 350 persons, in which each man is paired with each woman (but because men were not paired with other men and women were not paired with other women, this would not be round-robin).

Back, Schmukle, and Egloff (2010) studied what factors determined liking in one group of 73 first-year college students in a round-robin design.

Elfenbein, Curhan, Eisenkraft, Shirako, and Baccaro (2008) studied 149 persons in 31 4-or 5-person groups engaged in a series of negotiation tasks with multiple partners to examine whether there are consistent individual differences in negotiating skills.

In each of these cases, dyads are created from a group of persons and measurements are obtained from both members. In some cases, the group is an actual group as in Lam *et al.* (2011) and Back *et al.* (2010). In other cases, the group is just a nominal group, as in Finkel and Eastwick (2008) and Elfenbein *et al.* (2010). Also sometimes only a subset of all possible dyads in the group is formed. For instance, Finkel and Eastwick (2008) created only mixed-gendered dyads from each speed-dating event.

For such data structures, a very general model of nonindependence is the Social Relations Model (SRM). The SRM provides a general framework from which both social behavior and interpersonal perception can be studied. Here we provide a brief introduction to the model. For a more extended explanation of the SRM, see Back and Kenny (2010), and for fully detailed explanation, see Kenny and Kashy (2011) or Kenny (1994).

According to the SRM, a dyadic outcome measure can be broken into group-level effects, individual-level effects, and dyad-level effects. Consider an example of a team with six members, two of whom are Art and Beth. The researcher is interested in two variables: how helpful each team member perceives the other team members to be, and an observer's judgment of how cooperative each team member is with each other. Table 22.3 presents the SRM breakdown of these dyadic scores for both helpfulness and cooperativeness.

Table 22.3. Social Relations Model Components for Rating Measures and Behavioral Measures

Score	=	Group Mean	+	Art's Actor Effect	+	Beth's Partner Effect	+	Art's Relationship Effect with Beth
<i>Rating Measure</i> Art's perception of Beth's helpfulness	=	Group mean for helpfulness	+	Art's tendency to see all partners as helpful	+	Beth's tendency to be seen by all partners as helpful	+	Art's unique perception of Beth's helpfulness
<i>Behavioral Measure</i> Art's level of cooperativeness with Beth	=	Group mean for cooperativeness	+	Art's tendency to be cooperative with all partners	+	Beth's tendency to elicit cooperativeness from all partners	+	Art's unique amount of cooperativeness with Beth

According to the SRM, each dyadic score may be a function of four effects. First, at the group level, Art and Beth's team might on average have scored high on helpfulness relative to the other teams. That is, one component that contributes to Art's rating of Beth's helpfulness is the general level of helpfulness in the group as a whole: Some groups are more helpful than others. This first component is called the *group mean* and reflects the average level of the outcome score for the group, thereby modeling the group nonindependence in the data. Similarly, the degree to which Art cooperates with Beth in part reflects the group mean for cooperativeness, because some groups may cooperate more than others.

Next, at the individual level, Art's ratings or behavior may be consistent across all of his interactions with the other team members (Beth, Carol, Doug, Ed, and Felicia). For ratings of helpfulness, Art may tend to rate everyone as very helpful, and so part of Art's high rating of Beth's helpfulness may be a function of Art's general tendency to see others as helpful. In terms of cooperativeness, one factor that contributes to Art's level of cooperativeness with Beth is Art's general tendency to be cooperative with others. The tendency for a person to exhibit a consistent level of response across all interaction partners is generally called an *actor effect*. The meaning of the actor effect in the SRM differs from that for the APIM (see earlier discussion). In the APIM, the actor effect is the impact of a person's score on a predictor variable on that person's score on his or her outcome variable. In the SRM, the actor effect is the degree to which an individual provides consistent scores on the outcome variable across multiple dyads, there being no predictor variable.

The *partner effect* is also an individual-level effect, which measures the tendency for others to be consistent with a particular partner. Thus, for Art's rating of Beth's helpfulness the partner effect measures the tendency for Beth to be seen as helpful by all of her interaction partners. When outcome measures are

behavioral, the partner effect measures the degree to which certain individuals tend to elicit similar behavior from all of their interaction partners. In terms of cooperativeness, the partner effect measures the tendency for all group members to cooperate a great deal to Beth. As was true for the actor effect, the partner effect has different meanings for the APIM and the SRM. For the APIM, the partner effect is the degree to which a person's partner's score on a predictor affects the person's score on the outcome. For the SRM, the partner effect is the degree to which others behave in consistent ways on the outcome measure when interacting with a particular partner .

The terms “actor” and “partner” are generic terms, and other terms can and should be used in different contexts. For instance, for interpersonal perception, the terms “perceiver” and “target” are typically used. In nonverbal communication, the terms “decoder” and “encoder” might be used, whereas in aggression research, the terms “bully” and “victim” might be more appropriate. As we have noted, it is important not to confuse actor and partner in SRM with the terms in the APIM. In the SRM, actor and partner are random variables that explain variation across a person's dyadic relationships, whereas in the APIM, actor and partner refer to the effect of a given measured variable on another variable in one specific relationship. However in both cases, the actor effect refers to the influence of a variable on oneself, and the partner effect refers to the influence of a variable on the other person in the dyad.

The *relationship effect* is at the dyad level. For Art's rating of Beth's helpfulness, the relationship effect measures the degree to which Art sees Beth as especially helpful, over and above Art's general tendency to see others as helpful and over and above Beth's tendency to be seen by others as helpful. Thus, the relationship effect reflects the unique combination of two individuals after removing their individual-level tendencies as well as the group mean. For the cooperativeness variable, the relationship effect measures the degree to which Art is especially cooperative with Beth, after taking into account the group means and both Art's actor effect for cooperativeness and Beth's partner effect for cooperativeness.

The relationship effect can be separated from error only if there are multiple measures of the same underlying construct. For example, Art could rate Beth on two indicators of helpfulness, such as giving advice and listening. These two measures could be treated as indicators of helpfulness, and the part of Art's unique rating of Beth that is consistent across the two indicators would be treated as the relationship effect and any inconsistency across the two indicators

would be treated as error. Replications over time also may be used to partition relationship effects from error. That is, if Art and Beth interact two times, and cooperativeness is measured at each, Art's relationship effect for cooperativeness with Beth could be separated from noise attributable to chance fluctuations over time.

Table 22.4 depicts three simplified example data sets that contain only actor variance, only partner variance, and only relationship variance. The actor variance, measuring the degree to which some individuals see all partners as high on a trait and other individuals see all partners as low on a trait, is essentially (but not exactly) the variance among the row marginal means after averaging across partners. Thus, the actor variance is the row main effect. The partner variance, measuring the degree to which some individuals are seen by all actors as high on a trait and other individuals are seen by all actors as low on the trait, is essentially (but not exactly) the variance among the column marginal means after averaging across actors; it is the column main effect. The relationship variance, measuring the degree to which trait ratings are unique to particular pairings of actors and partners, is the variance attributable to the interaction between actor and partner. That is, the relationship variance is essentially (but not exactly) the variance in the cells, after the row marginal means (the main effect of actor) and the column marginal means (the main effect of partner) have been removed.

Table 22.4. *The Social Relations Model as a Two-Way Random Effects ANOVA*

Actor Variance Only						
		E	F	G	H	Row Marginal Means
Actor	A	1	1	1	1	1.0
	B	2	2	2	2	2.0
	C	3	3	3	3	3.0
	D	4	4	4	4	4.0
Column Marginal Means		2.5	2.5	2.5	2.5	
Partner Variance Only						
		E	F	G	H	Row Marginal Means
Actor	A	1	2	3	4	2.5
	B	1	2	3	4	2.5
	C	1	2	3	4	2.5
	D	1	2	3	4	2.5
Column Marginal Means		1.0	2.0	3.0	4.0	
Relationship Variance Only						
		E	F	G	H	Row Marginal Means
Actor	A	1	2	3	4	2.5
	B	2	3	4	1	2.5
	C	3	4	1	2	2.5
	D	4	1	2	3	2.5
Column Marginal Means		2.5	2.5	2.5	2.5	

The prototypical SRM is a round-robin in which every member of the group interacts with or rates every other individual in the group; the key requirement is that each dyadic combination provides an outcome score. Dyadic interactions can occur in the presence of the entire group (e.g., A, B, C, and D interact simultaneously), or they can occur one on one such that Art and Beth interact in one room while Carol and Doug interact in another, then A and C interact while B and D interact, and so on. Other designs, which are discussed in sources which we have cited earlier in the chapter, are possible. For instance, in a speed-dating study typically, people have dates only with partners of a different gender and so not all pairs are created. Typically, though not always, SRM data are reciprocal and so both persons in the dyad are measured.

As an example of the SRM, consider the recent analysis by Kenny and Livi (2009) who summarize seven SRM studies of leadership. In these studies, group members rated each other on multiple measures of leadership. They found that the dominant component, even larger than error variance, was partner variance, which was 42.7% of the total variance. Thus, group members agree who is the leader in the group and who is not. Relationship variance was nontrivial, accounting for 19.2% of the total variance. Interestingly, group accounted for virtually no variance, and so leadership is something that varies more within groups than between groups.

The SRM is a multilevel model, albeit a very complicated one. It may have as

many as five different levels: the observation, the measure, the dyad, the individual, and the group. The individual level contains variance attributable to actor and partner and those effects might well be correlated. Moreover, there may be interpersonal correlations between relationship effects (i.e., AB with BA).

Our description of the SRM has considered only random variables, that is, variances and covariance of random effects. Researchers often have an interest in the effects of experimental and individual difference variables, which are called *fixed* variables. This was the case for Lam *et al.* (2011) in their study examining the three-way interaction among cooperative goals, comparison to a higher performing team member, and future performance similarity to that member on harming behavior. Increasingly, we are seeing both the fixed and random effects of the SRM estimated by MLM programs (e.g., MLwiN in Lam *et al.*, 2011 and SAS in Kenny *et al.*, 2007). Old-fashioned analysis of variance programs (Kenny, 2012) and an R-based program TripleR (Schmukle, Schönbrodt, & Back, 2010) are also available. The reader should consult Biesanz (2010) for his version of the SRM to study accuracy in interpersonal perception.

It is not often realized that many social cognition studies are SRM studies. When researchers study the responses of participants to the same stimuli, who are persons, the participants take on the role of actor and the stimuli take on the role of partner. Obviously, the design is not reciprocal (the stimuli do not serve as participants), but the model is still the SRM. For example, the Correll, Park, Judd, and Wittenbrink (2002) “shooter” study contains participants (i.e., actors) who each view different stimuli (i.e., partners) who vary by both race and by whether they are holding a gun, these latter two variables being fixed variables. The outcome variable is the reaction time that it takes to make the determination of whether to shoot at the stimuli or not. Following Judd, Westfall, and Kenny (2012), it is essential that actor, partner, and relationship variance be estimated if we are to have a proper test of the theoretically important effect, the race by gun interaction.

Intergroup Research

Moreland, Hogg, and Hains (1994) found that approximately 38% of group research in social psychology is actually intergroup research, and we suspect that the percentage now is larger than 50. In an intergroup study, the group contains two subgroups, say A and B. For each person, members of their same group are called *in-group members* and members of the other group are called *outgroup*

members. Many intergroup studies are not interpersonal in the sense of this chapter because the groups are not small groups of actual others, but rather are social categories, for example, men and women or blacks and whites. Our discussion presumes that we have groups of individuals (e.g., groups of eight participants) that are divided into two subgroups (e.g., four As and four Bs). Sometimes the variable differentiating the subgroups is substantively meaningful (e.g., high versus low power) and other times it is not (e.g., over-versus underestimators).

We consider three different types of outcome variables from intergroup studies. The first type of outcome is one in which each individual makes a single rating or judgment, which we call a *single-measure outcome*. For instance, each member might be asked how much of a fixed reward should be given to his or her group. The second type of outcome is one in which each individual provides one in-group score and one outgroup score, which we call a *two-measure outcome*. For instance, a person may be asked how much he or she likes overall the in-group members and the outgroup members. Finally, the third type of measure is a *dyadic measure* in which a person rates every member of the in-group and of the outgroup. For instance, the person might be asked how much he or she likes every member of the group. We briefly consider the analysis of these three types of measures, with special emphasis on the different sorts of nonindependence in the data.

Single-measure outcomes. When each individual provides only a single score, the scores within each subgroup are nonindependent at the level of the group. That is, a person may be more similar (or possibly more different) to the members of his or her subgroup than he or she is to others in the study. This is essentially the same sort of nonindependence that we discussed earlier as group nonindependence. Note that this nonindependence can be computed separately for each of the two subgroups, for example, the A's and the B's. Moreover, the two subgroup effects may be correlated. For instance, if the outcome is how fair the members perceive a process to be, if the A's think that the process is fair, so might also the Bs.

Two-measure outcomes. In this case each individual responds to both the in-group and the outgroup. For each subgroup, there are two sorts of nonindependence, one for their rating of the in-group and one for the outgroup. Because there are two subgroups, there are four group effects, all of which may be correlated. For instance, if the subgroups are high and low status, there are four group effects: high-status in-group, high-status outgroup, low-status in-

group, and low-status outgroup. As was the case for the one-measure outcome, the group effects for a two-measure outcome can be correlated.

Dyadic outcomes. Dyadic outcomes with intergroup designs are very complicated. We use the study by Boldry and Kashy (1999) to briefly illustrate the different sources of nonindependence. For each in-group, we can compute group, actor, partner, and relationship variance. Moreover, for both types of outgroup ratings (the A's rating the B's and the B's rating the A's) a variance partitioning is possible. This type of design can be used to test the outgroup homogeneity hypothesis that predicts larger partner and relationship variances for in-group dyadic ratings than outgroup dyadic ratings. (As discussed in Boldry and Kashy, 1999, the outgroup homogeneity hypothesis is also supported when in-group actor variances are smaller than outgroup actor variances.)

In addition to estimating the variance in these effects, the different variance components can be correlated across in-group and outgroup ratings. For instance, the correlation between in-group actor effects and outgroup actor effects might measure whether individuals who perceive all of their in-group members positively tend to view the outgroup members positively as well. We urge the reader to consult Boldry and Kashy (1999) for a detailed illustration. Note finally that a heterosexual speed-dating study can be viewed as an “intergroup” dyadic study with only outgroup judgments. That is, men and women provide only ratings of outgroup members (i.e., rating of members of the other gender).

One-With-Many Design

A key construct in clinical psychology is working alliance, which includes the quality of the emotional bond between the client and the therapist, the level of collaboration in setting therapeutic goals, and the level of agreement on ways of achieving therapeutic goals (Bordin, 1979). Given that the therapeutic relationship is inherently a dyadic context, both the therapist and the client can report on their perceptions of alliance. If such a study was limited so that each therapist saw only a single patient, this would be a standard dyadic design and the APIM for distinguishable dyads might be an appropriate data analytic model. However, in many studies of alliance, each therapist sees multiple clients (e.g., Marcus, Kashy, & Baldwin, 2009; Manne, Kashy, Rubin, Hernandez, & Bergman, 2012), and so this design is actually what Kenny *et al.* (2006) call a one-with-many (OWM) design. Specifically, in an OWM design individuals are

linked together as members of a group because each person is tied to the same focal individual (e.g., the therapist). In this design the person who has multiple partners (the “one”) is designated as the *focal person* and the multiple others (the “many”) are designated as the *partners*. For instance, a manager (the focal person) may interact with many employees (partners), a teacher (the focal person) may interact with many students (partners), or a doctor (the focal person) may interact with many patients (partners). In this design there is assumed to be no direct relationship between the group members beyond their ties to the same focal person. Consider the following examples of the design taken from the literature:

An individual reports how jealous he or she felt in his or her four previous relationships (Hindy & Schwarz, 1994).

A total of 108 teachers’ expectancies of their 2,625 students’ performance are correlated with those students’ actual performance (Jussim & Eccles, 1992).

A person's three friends report on that person's personality (Vazire & Mehl, 2008).

Elderly persons rate their closeness and how frequently they interact with members of their social network (Cornwell, 2011).

A leader's liking is related to the quality of the leader-member exchange in 47, 4-person teams with 1 leader and 3 followers (Dockery & Steiner, 1990).

In each case a focal person judges or is judged by a set of partners with whom the focal person has some sort of relationship.

As an example, consider Dockery and Steiner's (1990) study of leader-member exchange, or LMX, in initial interactions. In this small-group study, one member was randomly assigned to the leader role and then the group members worked on a problem-solving task in which the leader's assignment was to guide the interaction so that the group members achieved consensus on the task. At the end of the interaction the leader reported on the quality of the LMX with each of the group members, and the group members also reported on their own LMX with the leader. As this example illustrates, the scores obtained in the OWM design can be provided by either the *one* (e.g., the leader's ratings of the members), the *many* (e.g., the members’ ratings of the leader), or both. When the data come from both the focal person and the partners, as in our example, it is called a *reciprocal* design.

Although the one-with-many design is often analyzed by treating partner as the unit of analysis, such an approach ignores the nonindependence attributable to the common focal person. Data from this design have a multilevel structure, with partner as the lower-level unit and focal person as the upper-level unit. The multilevel equation for partner j with focal person i is as follows:

$$Y_{ij} = (b_0 + b_{0i}) + b_1 X_{ij} + b_2 Z_j + e_{ij}$$

where X is a partner-level predictor variable and Z is a focal-person predictor variable. Nonindependence is modeled by the random component of the intercepts, d_{0i} . Note too that it is possible to treat a level 1 or focal person predictor (e.g., X) as a random variable and allow its effect (b_2) to vary across focal persons.

In the OWM design, interdependence occurs when scores for two partners who are tied to the same focal person are more similar to one another than are scores for partners who are tied to two different focal persons. However, the meaning of this interdependence depends on who generates the scores. If the data come from the focal person (i.e., all of the scores come from the leader), the nonindependence represents the consistency of the focal person's ratings. Here the nonindependence occurs because the focal person's ratings reflect his or her tendency to see all partners in a similar fashion. This effect is similar to the actor effect in the SRM (see earlier discussion). For instance, high levels of nonindependence in the leader example would indicate that some leaders tend to report high-quality exchanges with all group members whereas other leaders report low-quality exchanges with the other group members. Scores in this case also reflect factors specific to the unique combination of a particular partner with the focal person. In other words, the leader may rate exchanges with person A as high in quality, and this high rating may result from two elements: the leader tends to evaluate everyone's exchange positively (the actor effect previously discussed), and the leader thinks that the exchange with A was exceptionally positive.

In contrast, if the data come from the many (i.e., each of the partners), the nonindependence represents a consensus among the partners concerning their perceptions of the focal person. In the example, each group member rates the quality of his or her exchange with the leader, and so the nonindependence occurs because the members' ratings of exchange with the leader may be similar. When multiple perceivers report on their perceptions of the same target, we can

estimate an effect similar to the partner effect in the SRM. When there is large variation in these partner effects, the implication is that some leaders are seen as establishing strong exchanges by all of their group members, but other leaders' exchanges are rated as inferior by all of their group members. In this case, each score for a partner rating the focal person may reflect both the partner effect as well as a unique component. For example, person A may view his leader as establishing high quality exchanges whereas other group members may report more negative exchanges with the leader.

In the reciprocal OWM design both the focal person and the partners provide scores, and so this design allows for estimation of both actor and partner effects. That is, in a single study we can estimate the tendency for a leader to report positive exchanges with all group members, and we can also estimate the tendency of the group members to agree in their reports of exchange quality with the leader. Thus, in both cases the focus is on the focal person, and so the reciprocal design allows researchers to examine the degree to which these two effects correspond with one another. In the example, if the correspondence is positive, it would indicate that leaders who generally report positive exchanges with other group members are generally viewed as establishing positive exchanges by those group members. This correspondence has been called *generalized reciprocity*, and it is estimated as the correlation between the actor and partner effects.

The reciprocal design also allows us to examine dyad-level correspondence in ratings. *Dyadic reciprocity* measures the correspondence between the two "errors" in the reciprocal design: If the leader sees exchange with A as especially positive (more so than other members), does A see the exchange as especially strong (again, more so than other workers)? Dyadic reciprocity often occurs on affective measures such as liking. A detailed discussion of the OWM design and its analysis using SPSS can be found in Marcus et al., (2009) and Kashy and Donnellan (2012). One can also consult Kenny *et al.* (2010) for an illustration of agreement about the communication in doctor-patient dyads.

Although the OWM design allows for estimation of several effects, it has some limitations. Continuing with the example, because each group member is tied to only one leader, the member's positive perception of exchange with the leader may reflect the fact that he or she truly thinks that the particular leader did well, but it might also reflect the fact that the worker is a positive person who would think that almost any leader did well. To separate these factors, each individual must participate in multiple dyads – which would then be an SRM

design.

Finally, if one has fixed variables in the OWM design, they can be included in the model. A fixed variable (e.g., intelligence) can be measured only for the partners, only for the focal persons, or for both. If it is measured for both, then the investigator has a choice. Considering the variable gender, there are two ways gender can be coded: one way is to use the gender of the focal person and gender of the partner, and the other way is to use the gender of the actor (i.e., the person providing the response) and the gender of the partner. The latter choice results in an APIM analysis of OWM data.

Conclusion

This chapter presents research designs and data analytic methods that recognize that individuals within a group or dyad may influence one another. We have described how the degree of interrelatedness of scores from dyads and groups can be assessed. We have also described a number of data analytic approaches that, rather than assuming independence, instead explicitly model interdependence.

This chapter is merely an introduction into the analysis of nonindependent data in dyads and groups. There are several other important methods that have not been discussed. Social network analyses, for example, are a set of techniques that can be used to model the interrelatedness of individuals within a social group (Wasserman & Faust, 1994). Nor have we discussed over-time dyadic (Bolger & Laurenceau, 2013) and group data. There are also specialized designs in which individuals are members of more than one group (Kenny, Hallmark, Sullivan, & Kashy, 1993), as well as models for triadic processes (Bond, Horn, & Kenny, 1997).

As we noted in the introduction of this chapter, if social psychology is to grow and develop as the study of truly social behavior, researchers need to become proficient in the kinds of designs and analyses that allow them to study individuals within actual social contexts.

References

- Ackerman, R. A., Kashy, D. A., Donnellan, M. B., & Conger, R. D. (2011). Positive engagement in family interactions: A social relations perspective. *Journal of Family Psychology*, 25, 719–730.

- Anderson, C. A., Buckley, K. E., & Carnagey, N. L. (2008). Creating your own hostile environment: A laboratory examination of trait aggressiveness and the violence escalation cycle. *Personality and Social Psychology Bulletin*, 34, 462–474.
- Anderson, L. R., & Ager, J. W. (1978). Analysis of variance in small group research. *Personality and Social Psychology Bulletin*, 4, 341–345.
- Back, M. D., & Kenny, D. A. (2010). The social relations model: How to understand dyadic processes. *Social and Personality Psychology Compass*, 4, 855–870.
- Back, M. D., Schmukle, S. C., & Egloff, B. (2010). Why are narcissists so charming at first sight? Decoding the narcissism – popularity link at zero acquaintance. *Journal of Personality and Social Psychology*, 98, 132–145.
- Biesanz, J. C. (2010). The social accuracy model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate Behavioral Research*, 45, 853–885.
- Boldry, J. G., & Kashy, D. A. (1999). Intergroup perception in naturally occurring groups of differential status: A social relations perspective. *Journal of Personality and Social Psychology*, 77, 1200–1212.
- Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York: Guilford Press.
- Bond, Jr., C. F., & Kenny, D. A. (2002). The triangle of interpersonal models. *Journal of Personality and Social Psychology*, 83, 355–366.
- Bond, C. F., Horn, E. M., & Kenny, D. A. (1997). A model for triadic relations. *Psychological Methods*, 2, 79–94.
- Bonito, J. A. (2002). The analysis of participation in small groups. Methodological and conceptual issues related to interdependence. *Small Group Research*, 33, 412–438.
- Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of working alliance. *Psychotherapy: Theory, Research, and Practice*, 16, 252–260.
- Cillessen, A. H. N., Jiang, X. L., West, T. V., & Lazkowski, D. K. (2005).

- Predictors of dyadic friendship quality in adolescence. *International Journal of Behavioral Development*, 29, 165–172.
- Cornwell, B. (2011). Independence through social networks: Bridging potential among older women and men. *The Journals of Gerontology B*, 66, 782–94.
- Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, 83, 1314–1329.
- Dabbs, Jr., J. M., & Ruback, R. B. (1987). Dimensions of group process: Amount and structure of vocal interaction. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 20, pp. 123–169). San Diego, CA: Academic Press.
- Dockery, T. M., & Steiner, D. D. (1990). The role of the initial interaction in leader-member exchange. *Group & Organization Studies*, 15, 395–413.
- Elfenbein, H. A., Curhan, J. R., Eisenkraft, N., Shirako, A., & Baccaro, L. (2008). Are some negotiators better than others? Individual differences in bargaining outcomes. *Journal of Research in Personality*, 42, 1463–1475.
- Finkel, E. J., & Eastwick, P. W. (2008). Speed-dating. *Current Directions in Psychological Science*, 17, 193–197.
- Garcia, R. L., Kenny, D. A., & Ledermann, T. (2013). *Moderation in the actor-partner interdependence model*. Unpublished paper, Princeton University.
- Goldberg, A. E., Smith, J. Z., & Kashy, D. A. (2010). Pre-adoptive factors predicting lesbian, gay, and heterosexual couples' relationship quality across the transition to adoptive parenthood. *Journal of Family Psychology*, 24, 221–232.
- Hindy, C. G., & Schwarz, J. C. (1994). Anxious romantic attachment in adult relationships. In M. B. Sperling & W. H. Berman (Eds.), *Attachment in adults: Clinical and developmental perspectives* (pp. 179–203). New York: Guilford Press.
- Hox, J. (2010). *Multilevel analyses: Techniques and applications*, 2nd ed. Mahwah, NJ: Erlbaum.
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive

but largely ignored problem. *Journal of Personality and Social Psychology*, 103, 54–69.

Jussim, L., & Eccles, J. (1992). Teacher expectations: II. Construction and reflection of student achievement. *Journal of Personality & Social Psychology*, 63, 947–961.

Kashy, D. A., & Donnellan, M. B. (2012). Conceptual and methodological issues in the analysis of data from dyads and groups. In K. Deaux & M. Snyder (Eds.), *The Oxford handbook of personality and social psychology* (pp. 209–238). New York: Oxford University Press.

Kashy, D. A., & Kenny, D. A. (2000). The analysis of data from dyads and groups. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 451–477). New York: Cambridge University Press.

Kashy, D. A., & Kenny, D. A. (2012). Test of distinguishability using multilevel modeling. Retrieved November 2012 from http://www.davidakenny.net/doc/indistinguishability_mlm.pdf.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford Press.

Kenny, D. A. (2012). SRM software. Retrieved November 2012 from <http://davidakenny.net/srm/srmp.htm>.

Kenny, D. A., & Cook, W. L. (1999). Partner effects in relationship research: Conceptual issues, analytic difficulties, and illustrations. *Personal Relationships*, 6, 433–448.

Kenny, D. A., & Garcia, R. L. (2012). Using the Actor-Partner Interdependence Model to study the effects of group composition. *Small Group Research*, 43, 468–496.

Kenny, D. A., Hallmark, B. W., Sullivan, P., & Kashy, D. A. (1993). The analysis of designs in which individuals are in more than one group. *British Journal of Social Psychology*, 32, 173–190.

Kenny, D. A., & Kashy, D. A. (2011). Dyadic data analysis using multilevel modeling. In J. Hox & J. K. Roberts (Eds.), *The handbook of advanced multilevel analysis* (pp. 335–370). London: Taylor & Francis.

- Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York: Guilford Press.
- Kenny, D. A., & La Voie, L. (1985). Separating individual and group effects. *Journal of Personality and Social Psychology*, 48, 339–348.
- Kenny, D. A., & Ledermann, T. (2010). Detecting, measuring, and testing dyadic patterns in the Actor-Partner Interdependence Model. *Journal of Family Psychology*, 24, 359–366.
- Kenny, D. A., & Livi, S. (2009). A componential analysis of leadership using the Social Relations Model. In F. J. Yammarino & F. Dansereau (Eds.), *Multilevel issues in organizational behavior and leadership* (Vol. 8, pp. 147–191). Bingley, UK: Emerald.
- Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology*, 83, 126–137.
- Kenny, D. A., Veldhuijzen, W., Weijden, T., Leblanc, A., Lockyer, J., Légaré, F., & Campbell, C. (2010). Interpersonal perception in the context of doctor-patient relationships: A dyadic analysis of doctor-patient communication. *Social Science and Medicine*, 70, 763–768.
- Kenny, D. A., West, T. V., Cillessen, A. H. N., Coie, J. D., Dodge, K. A., Hubbard, J. A., & Schwartz, D. (2007). Accuracy in judgments of aggressiveness. *Personality and Social Psychology Bulletin*, 33, 1225–1236.
- Kivlighan, D. R., Kivlighan, D., & Cole, O. (2012). The group's absence norm and commitment to the group as predictors of group member absence in the next session: An actor-partner analysis. *Journal of Counseling Psychology*, 59, 41–49.
- Klumb, P., Hoppmann, C., & Staats, M., (2006). Work hours affect spouse's cortisol secretion-for better and for worse. *Psychosomatic Medicine*, 68, 742–746.
- Lam, C., Van der Vegt, G. S., Walter, F., & Huang, X. (2011). Harming high performers: Social comparison and interpersonal harming in work teams. *Journal of Applied Psychology*, 96, 588–601.
- Ledermann, T., & Kenny, D. A. (2012). The common fate model for dyadic data: Variations of a theoretically important but underutilized model. *Journal*

of Family Psychology, 26, 140–148.

- Ledermann, T., Macho, S., & Kenny, D. A. (2011). Assessing mediation in dyadic data using the Actor-Partner Interdependence Model. *Structural Equation Modeling*, 18, 595–612.
- Lemay, Jr., E. P., & Clark, M. S. (2008). How the head liberates the heart: Projection of communal responsiveness guides relationship promotion. *Journal of Personality and Social Psychology*, 94, 647–671.
- Manne, S., Kashy, D. A., Rubin, S., Hernandez, E., & Bergman, C. (2013). Therapist and patient perceptions of alliance and progress in psychological therapy for women diagnosed with gynecological cancers. *Journal of Consulting and Clinical Psychology*, 80, 800–810.
- Marcus, D. K., Kashy, D. A., & Baldwin, S. A. (2009). Studying psychotherapy using the one-with-many design: The therapeutic alliance as an exemplar. *Journal of Counseling Psychology*, 56, 537–548.
- Moreland, R. L. (2010). Are dyads really groups? *Small Group Research*, 41, 251–267.
- Moreland, R. M., Hogg, M. A., & Hains S. C. (1994). Back to the future: Social psychological research on groups. *Journal of Experimental Social Psychology*, 30, 527–555.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.
- Riggs, S. A., Cusimano, A. M., & Benson, K. M. (2011). Childhood emotional abuse and attachment processes in the dyadic adjustment of dating couples. *Journal of Counseling Psychology*, 58, 126–138.
- Schmukle, S. C., Schönbrodt, F. D., & Back, M. D. (2010). *TripleR: A package for round robin analyses using R (version 0.5.3)*. Retrieved November 2012 from <http://www.persoc.net/ToolBox/TripleR>.
- Snijders, T. A., & Bosker, R. J. (1999) *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.
- Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: The accuracy and unique predictive validity of self-and other ratings of daily behavior. *Journal of Personality and Social Psychology*, 95, 1202–1216.

Wasserman S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Webster, G. D., Brunell, A. B., & Pilkington, C. J. (2009). Individual differences in men's and women's warmth and disclosure differentially moderate couples' reciprocity in conversational disclosure. *Personality and Individual Differences*, 46, 292–297.

Weingart, L. R., Brett, J. M., Olekalns, M., & Smith, P. L. (2007). Conflicting social motives in negotiating groups. *Journal of Personality & Social Psychology*, 93, 994–1010.

¹ In fact, if either the estimate of the variance of f or d is small, it is advisable to reestimate the model setting the term to zero.

* Parts of this chapter were adapted from a prior version (Kashy & Kenny, 2000).

Chapter twenty-three Nasty Data

Unruly, Ill-Mannered Observations Can Ruin Your Analysis

Gary H. McClelland*

Researchers confronting their own data often find those data to be more unruly, ill-mannered, and irascible than the well-behaved, cooperative data found in textbook examples. Irascible data that slap us in the face at least get our attention. More dangerous are those stealthy, sinister observations that can go undetected and yet have a disproportionate and untoward effect on our analyses. This chapter describes techniques for detecting and taming those nasty data that otherwise could ruin your analyses.

Source of Problems

The supposedly tried-and-true statistical methods can be remarkably fragile in some situations. In particular, those methods that minimize sums of squares – including the most common statistical procedures used by social psychologists such as *t*-tests, ANOVAs, and regression – are very sensitive to outliers. Estimates of group means or regression parameters can sometimes be dramatically affected by just a single outlier observation. Most textbooks from which social psychologists learn their statistics indicate that the standard procedures assume normality and homogeneity of variance (equal variances within each group or equal variances around each predicted criterion in regression). Tables of critical values or those *p*-values produced by statistical programs as well as statistical power analysis depend strongly on those assumptions being correct. But few textbooks provide information about how to detect violations of the normality and homogeneity of variance assumptions in one's data or what to do about them when they are detected.

The good news is that the standard statistical methods are remarkably robust against some violations of the assumptions. However, there are some violations of those assumptions that can wreak havoc on one's analysis. In this chapter we give advice about how to separate the benign from the killer violations.

There are many ways in which one's data can become nasty and ill-mannered.

Data recording and data entry errors have the potential for introducing whopping outliers. We may believe we are sampling from what appears to us to be one population but which in fact is several distinct populations, one possibly much smaller than the others. Or nature may just throw us a curveball now and then to keep the game interesting. In whatever ways our data become irascible, it is important for us to detect the potential problems.

Most statistical analyses are performed by computers, but unfortunately the standard output from many programs just provides the usual test statistics, intermediate steps in the calculation of those test statistics, and their associated probabilities under the null hypothesis. Such outputs seldom provide clues about nasty data that might be lurking in the analysis. However, almost all computer packages also now provide adequate graphical and statistical procedures for detecting outliers and data that violate the standard assumptions. And the relevant procedures are mostly graphical and fairly simple, so there is no longer any excuse for not examining one's data for outliers and other ill-mannered data.

Why Worry?

Many psychologists, especially those far removed from their last graduate school statistics course, believe that the standard least-squares statistical procedures are relatively robust against all but the most serious kinds of nasty data. Indeed, early Monte Carlo studies (in which data were randomly sampled so as to violate specific assumptions) did indicate that basic tests such as Student's *t*-test were generally robust against certain violations. However, these studies have been greatly overgeneralized, and more recent theoretical and Monte Carlo studies have identified some unruly types of data – data not that uncommon in real research – that can have disastrous effects on analyses.

The old claims for robustness found in many statistics textbooks written for psychologists should be discounted for three reasons. First, the early robustness studies often considered quite restricted violations. For example, the frequently cited study by Box (1954) of the effects of unequal variances on the Type I level of Student's *t*-test considered ratios of standard deviations no more extreme than $\sqrt{3}$ to 1. However, subsequent research has shown that Student's *t*-test is remarkably sensitive when standard deviation ratios are greater than that, especially if sample sizes are unequal. Wilcox (1996) gave this example: If the two group sizes are 21 and 41 and the standard deviation of the smaller group is four times that of the larger (Wilcox, 1987 demonstrates that such extreme ratios are not uncommon in the social sciences), then the putative $\alpha = 0.05$ Student's *t*-

test actually has a Type I error rate of approximately 0.15.

A second reason for discounting the old claims for robustness is that robustness with respect to one issue, such as Type I error rate, does not imply robustness with respect to other important statistical issues. In particular, many test statistics whose Type I error rates are not appreciably altered by assumption violations often have greatly diminished statistical power. Wilcox (1996, p. 114) presents an example where a “contaminated” normal distribution (a small percentage of the cases are sampled from a normal distribution with a much larger standard deviation but the same mean) reduces the power of Student's *t*-test from .99 to .33. Furthermore, there are few robustness studies of other statistical measures such as confidence intervals and effect sizes that are increasingly recommended for psychological research.

A third reason for ignoring robustness claims is that even if the inferential test statistics were robust against all assumption violations, the tools described here for assessing assumption violations would still be valuable for providing researchers with a better understanding of their data. Instead of just using statistical procedures canonically to “bless” one's data to the editor's satisfaction, the goal of statistical analysis should be to understand one's data better. We cannot identify nasty, inconsistent data unless we have successfully modeled most of our data, and in turn identifying inconsistencies often shows the way to a more complete model of the data. Velleman (1997) reminds us that this is not a new idea with this quotation from Bacon's *Novum Organum*: “[E]rrors of Nature, sports and monsters correct the understanding in regard to ordinary things, and reveal general forms. For whoever knows the ways of Nature will more easily notice her deviations; and, on the other hand, whoever knows her deviations will more accurately describe her ways” (II 29).

This chapter is devoted to methods for knowing the deviations so that we may more accurately describe nature's ways. For didactic reasons, I first consider detection or diagnosis techniques, and then I suggest some useful remedies. However, the reader should be aware that in practice, data analysis using these techniques is highly iterative, with back and forth testing, using the same detection techniques, to assess whether the remedies have been successful.

Methods for Detecting Nasty and Ill-Mannered Data

The methods for detecting nasty and ill-mannered data depend somewhat on the type of analysis. However, most of the ideas are common to all the methods. We

begin by considering data from a single group, move on to comparisons among groups, and conclude with simple and multiple regression. In social psychology we seldom test statistical hypotheses with a single group because we seldom have adequate a priori null hypotheses to justify such tests. Nevertheless, examining the data for a single group is a good place to start because it provides a simple context in which to introduce several of the primary methods of detecting outliers and assumption violations.

Single Groups

Normality Assumption.

When applying standard least-squares statistical techniques to data from a single group, or to data from a single cell from a factorial design, we assume that the data are sampled from a normal distribution, such as the smooth curve depicted in [Figure 23.1](#). However, the smooth density curve of [Figure 23.1](#) only emerges if there are many, many observations. With small samples such as those commonly used in social psychology experiments and even with a couple of hundred observations in a field study, the empirical density function or histogram seldom closely approximates the normal density curve, even if the data are from a normal distribution. For example, the following 25 data values were sampled randomly from a normal distribution with mean 50 and standard deviation 10:

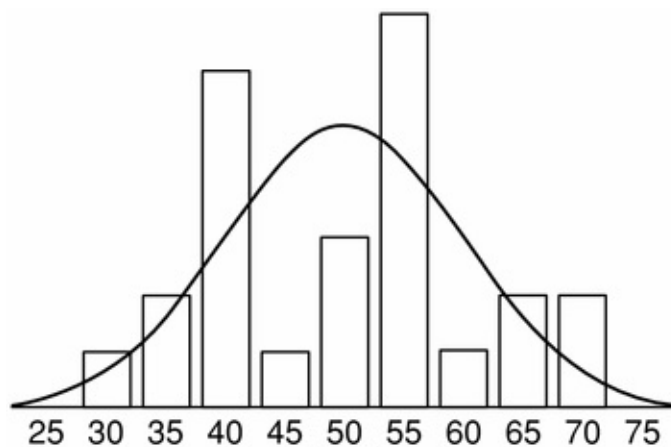


Figure 23.1. Density function for the normal distribution superimposed on histogram of 25 random observations from normal distribution.

38 61 42 57 70 52 51 29 51 34 38 72 55 55 54 39 54 44 57 55 40 42 63 65
36.

The histogram for these data, also in [Figure 23.1](#), shows considerable lumpiness. Using only the histogram, judging whether these data are from a normal distribution is difficult.

There are statistical tests for normality (Madansky, 1988, pp. 20–53, thoroughly describes the best methods). However, both these tests and viewing the histogram focus our attention on many types of nonnormality that are not problematic. [Figure 23.2](#) displays a Cauchy distribution (which results from the ratio of two normally distributed variables) on the left and a uniform distribution on the right. Both visual inspection and the normality tests (if we performed them) easily reveal that the uniform distribution is not normal, but the case of the Cauchy distribution is not so clear. However, in terms of the effect on a least-squares analysis, the Cauchy distribution is by far the more dangerous, whereas the uniform distribution poses virtually no problems. Extreme observations in the tails of the distribution pose the greatest danger of overwhelming the other observations and distorting the analysis. The “squares” in least-squares analysis, inherent in ANOVA and regression, implicitly give extra weight to those extreme observations in the tails; thus it is those observations that can most seriously distort a least-squares analysis. Although it is difficult to see in [Figure 23.2](#), the tails of the Cauchy distribution are much thicker (that is, extreme observations are relatively more likely) than those of the normal distribution. In contrast, the tails of the uniform distribution are in a sense chopped off, so there is little threat to the analysis. Although eyeballing the empirical density or using the statistical tests of normality considers the whole distribution, it is really only the tails that require our close attention.

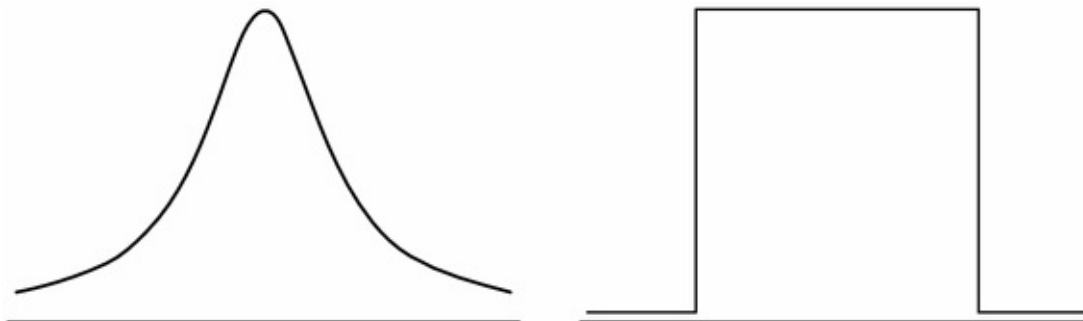


Figure 23.2. Examples of two nonnormal distributions: Cauchy (left) and uniform (right).

Fortunately, there are better and fairly simple ways to examine the tails of the empirical distribution. The normal quantile-quantile (q-q) plot (Chambers, Cleveland, Kleiner, & Tukey, 1983; Wilk & Gnanadesikan, 1968) is one such graphical method for examining the tails. Normal q-q plots were popular when normal probability graph paper was available, and have regained favor as they have become increasingly available in computer statistical packages. They are easy enough to construct directly and doing so makes clear what the computer packages do when they generate normal q-q plots. The idea is simple: We plot the empirical quantiles (the ordered data) against quantiles (i.e., z-scores) for the normal distribution. For example, one point in the normal q-q plot is the median of the data and the z-score = 0, which corresponds to the median of a normal distribution. In Table 23.1, the first column is the ordered data from the example of Figure 23.1, and i is just the index of the data. The next column is the fraction f_i of the observations that are equal to or less than the data value for the given row. We assume that each data value is in the middle of a histogram category by defining $f_i = (i - .5)/n$. Subtracting .5 avoids problems at the ends and ensures, for example, that for the median $f = .5$ (other definitions sometimes used for f_i to correct for problems at the ends are $f_i = i/[n + 1]$ and $f_i = [i - 3/8]/[n + 1/4]$). The last column in Table 23.1 is the z-score corresponding to each fraction f_i .

Table 23.1. Empirical and Normal Quantiles for Sample Data

Empirical			Normal
Quantile			Quantile
Data	i	f_i	$z(f_i)$
29	1	0.02	-2.05
34	2	0.06	-1.55
36	3	0.10	-1.28
38	4	0.14	-1.08
38	5	0.18	-0.92

39	6	0.22	-0.77
40	7	0.26	-0.64
42	8	0.30	-0.52
42	9	0.34	-0.41
44	10	0.38	-0.31
51	11	0.42	-0.20
51	12	0.46	-0.10
52	13	0.50	0.00
54	14	0.54	0.10
54	15	0.58	0.20
55	16	0.62	0.31
55	17	0.66	0.41
55	18	0.70	0.52
57	19	0.74	0.64
57	20	0.78	0.77
61	21	0.82	0.92
63	22	0.86	1.08
65	23	0.90	1.28
70	24	0.94	1.55
72	25	0.98	2.05

The normal q-q plot is simply a scatterplot of data against $z(f_i)$ as in [Figure 23.3](#). If the data are from a normal distribution with mean μ , and standard deviation σ , then data can be described by the following straight line:

$$\text{data} = \mu + \sigma z(f_i). \quad (23.1)$$

In other words, if the data are all telling more or less the same story, then the points should fall on a line with intercept equal to the mean and slope equal to the standard deviation. Points that deviate from the line, especially at either end, either because they are outliers or because the data distribution is skewed, are trying to tell a different story. In those cases, the one story told by the mean and the standard deviation does not apply to all the data points. In the example of [Figure 23.3](#), however, the data do fall closely to the line representing a normal distribution with intercept equal to the sample mean of 50.2 and a slope equal to the sample standard deviation of 11.5. Hence, these data are much more consistent with a normal distribution than implied by inspection of the lumpy histogram in [Figure 23.1](#). (Note that Shapiro & Wilk's [1968] W , one of the more popular statistical tests of normality, simply formalizes this regression concept in a manner accounting for the inherent nonindependence among the estimated residuals.)

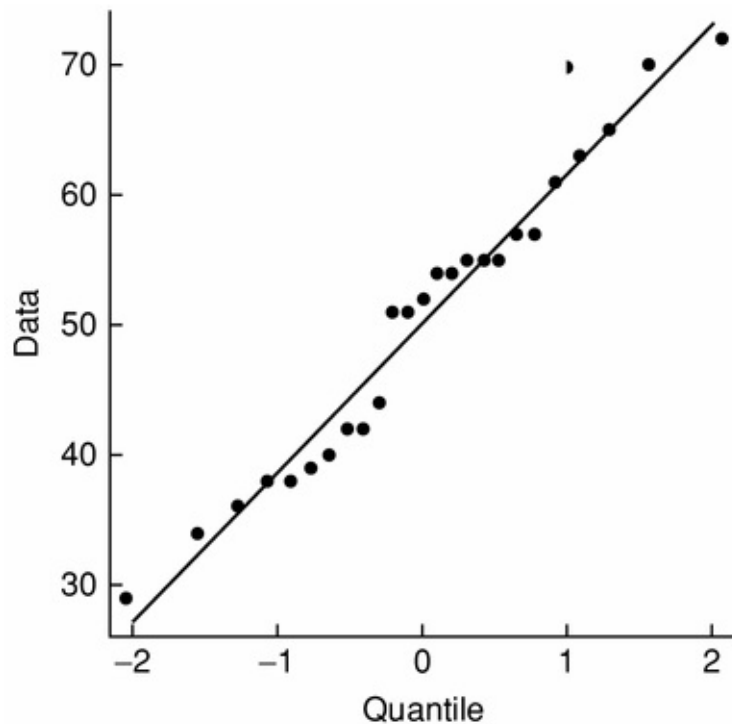


Figure 23.3. Normal quantile-quantile plot for sample data from a normal distribution.

Figure 23.4 shows four normal q-q plots. Those in the top row – data from normal and uniform distributions, respectively – are not threatening to a least-squares analysis because they indicate no more extreme observations in the tails than we would expect for a normal distribution. However, those in the bottom row – data from Cauchy and chi-square distributions, respectively – are very dangerous because they indicate the observations in the tails are more extreme than expected for a normal distribution. The plot for the normal distribution shows the linear pattern expected in a q-q plot. The plot for the uniform distribution is essentially linear in most of the middle range but, for at least one end, the slope flattens. Flat slopes at the ends imply that the tails of the distribution are thinner than those of a normal distribution. Thin tails generally are not problematic because they imply that unruly extreme observations, to which least-squares techniques are very sensitive, are less likely to occur than they would if normality were satisfied. In contrast, the plot for the Cauchy distribution illustrates a distribution with much thicker tails than the normal; the Cauchy plot is also essentially linear in most of the middle range but at either end the slope dramatically steepens. The data values on the steep part of the q-q plot are effectively outliers that have a disproportionate effect on a least-squares analysis. Finally, the q-q plot for the chi-square distribution illustrates the

appearance of an asymmetric or skewed distribution; there are essentially two slopes: a flat slope below the median, indicating a thin tail, and a very steep slope above the median, indicating a thick tail. Skewed distributions can also have adverse effects on standard statistical procedures. Unless there is a steep slope at either end of the normal q-q plot, then you probably do not need to worry about violations of the normality assumption. In other words, thin-tail violations of normality are generally benign, whereas thick-tail violations can kill an analysis.

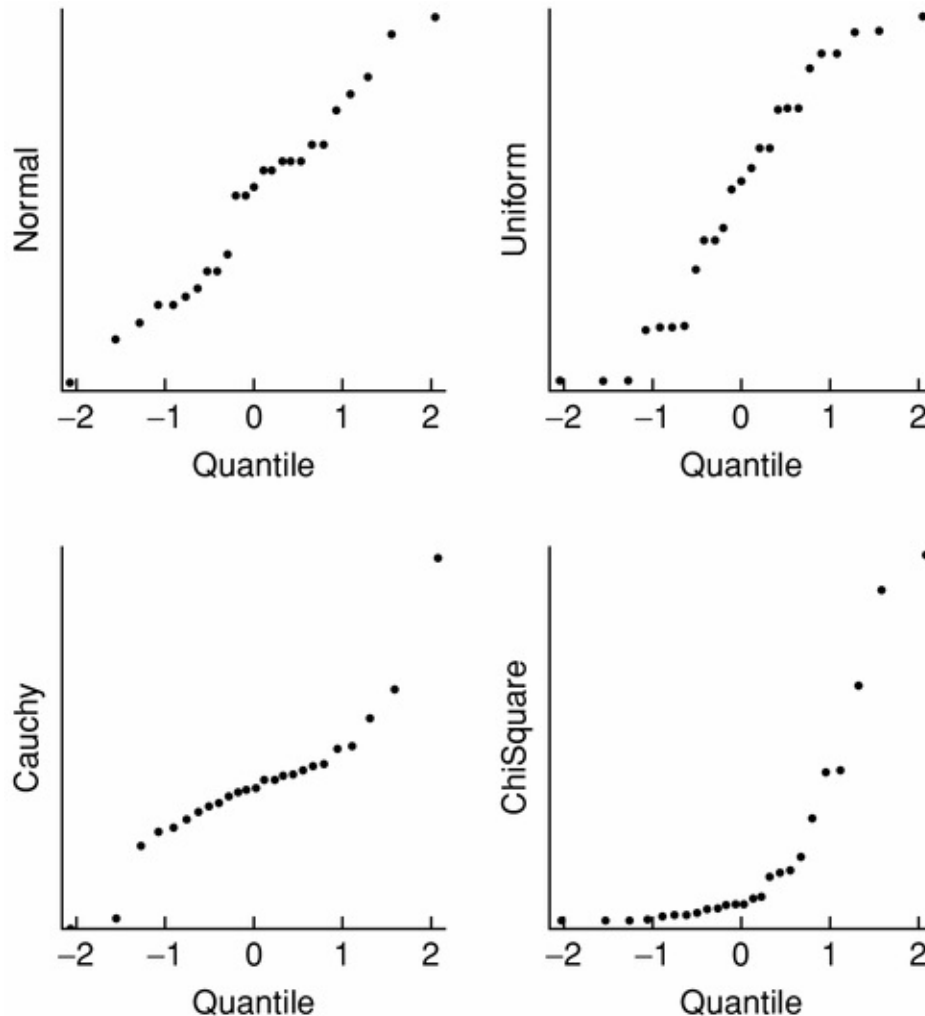


Figure 23.4. Normal q-q plots for four distributions: (clockwise from upper left) normal, uniform, Cauchy, chi-square.

More and more software packages provide normal q-q plots. Although most plot the empirical data against the equivalent z-score quantiles as illustrated in [Figures 23.3](#) and 23.4, some reverse the axes. For reversed axes, exchange “flat” for “steep” in the aforementioned descriptions. Also, some programs normalize

the data on the vertical axis so that the slope of the line is necessarily one. Variants of normal q-q plots also exist. For example, Atkinson (1981, 1982) argues that half-normal plots, in which the absolute value of the studentized deleted residual (see discussion later in the chapter) is plotted against the normal quantiles increases diagnosticity for small and moderate samples (say, less than 60 observations).

Outliers.

If the data really are from a single group, then they all should more or less be telling the same story, especially the story about what the typical value is for the group. If one or a few observations are telling a very different story from the other observations, then the assumption that the data come from a single group ought to be questioned. Outliers often identify themselves by causing steep tails in the normal q-q plots considered earlier. There are also statistically principled methods for identifying outliers. One method is to compute the standardized residual for each observation; that is,

$$r'_i = \frac{Y_i - \bar{Y}}{s} \sqrt{\frac{n}{n-1}}. \quad (23.2)$$

Standardized residuals are often compared with z-scores from normal distributions. With the sample sizes used in social psychology experiments or even surveys, z-scores greater in absolute value than about 2.6 should be rare. For example, if the underlying distribution is normal, then only about 1 out of every 100 observations should have a z-score greater than 2.6. For 1,000 observations, only about 4 observations should have absolute z-scores greater than 3. One should clearly be very suspicious of any z-scores greater than 4 or 5.

One problem with using z-scores is that an outlier may distort the estimated mean and standard deviation so that the outlier no longer looks extreme. A solution is to leave out an observation, recalculate the mean and standard deviation of the remaining observations, and then calculate the r-score. Actually, it is more appropriate to compare the resulting number to Student's *t*-distribution, so we have

$$r_i^* = \frac{Y - \bar{Y}_{[i]}}{s_{[i]}} \quad (23.3)$$

is distributed as Student's t -distribution with $df = n - 2$, where the $[i]$ as a subscript indicates that the i th observation has been omitted from calculating the mean and standard deviation. This value is exactly the same as the t -statistic for testing the coefficient for a dummy code ($= 1$ for the i th observation and $= 0$ for all other observations) onto which Y was regressed. This value of t is also sometimes known as the *studentized deleted residual*. It would be tiresome to recompute n means and standard deviations, each time leaving out one observation. Fortunately, this is not necessary because, as shown by Atkinson (1985):

$$r_i^* = r_i' \sqrt{\frac{n - 1 - r_i'^2}{n - 2}}. \quad (23.4)$$

Additionally, many computer packages now compute studentized deleted residuals. SAS (1989) and many other programs refer to this outlier index as RSTUDENT.

Sometimes there are a priori suspicions that an observation might be an outlier. For example, one might suspect the data from a participant who completed a questionnaire in 10 minutes that all other participants required more than an hour to complete. Or one might be wary of data from a participant whose experimental session was interrupted by a fire alarm. If so, then it is appropriate to compare the studentized deleted residual for that observation against Student's t -distribution in a test of whether one can reject the null hypothesis that the observation is telling the same story as all the other observations.

Without a priori doubts about a few observations, it is still appropriate to use the studentized deleted residual to screen for outliers. However, computing the studentized deleted residual for all the observations increases the likelihood of Type I errors. Hence, it is prudent to use the Bonferroni adjustment – using α/n instead of α – when identifying statistically significant outliers. For 100 observations, any observation with a studentized deleted residual greater in absolute value than 3.6 is significantly different ($p < .05$) from the other observations; for 1,000 observations, the critical value is 4.07. Thus, without worrying about tables of critical values and Bonferroni adjustments, any studentized deleted residual greater than 4 should be considered as an outlier.

Two Groups

A very common statistical test is the comparison of two groups such as control versus treatment or the comparison of two demographically defined groups such as men versus women. The assumptions underlying the standard statistical procedures such as Student's t -test are that the data from each group come from the same normal distribution except for a possible shift in location (i.e., the difference in the means). In particular, the variance of the normal distribution must be the same within each group. The procedures considered earlier for single groups can of course be applied to the data from each individual group. Rather than making two separate q-q plots, it is more common to plot the errors – the deviations from the within-group means – in a single q-q plot. Such a plot assesses the normality assumption underlying Student's t -test for independent groups and can reveal outliers, but it does not assess the assumption that the variances are the same within each group (homogeneity of variance). The q-q plot of the errors should fall on a line having a mean of zero and a slope equal to the estimated common standard deviation (i.e., the root mean squared error).

Homogeneity of Variance.

Just like testing for normality, there are formal statistical procedures for testing homogeneity of variance (see Madansky, 1988 for a review of many such tests). However, these tests themselves are often not robust to violations of normality and the presence of outliers. Furthermore, even when they detect violations, they do not provide any diagnosis that might lead to a remedy. There are, fortunately, several graphical techniques that are both useful for detecting violations of homogeneity of variance and for assessing the cause of the violations. The mean and median coincide for a normal distribution so the difference in the 50th percentiles estimates the difference in the means. Not as well known is that equivalent quantiles, as illustrated in the left column of Figure 23.5, from two identical normal distributions are also the same distance apart and so estimate the difference in the means. If the distributions for the two groups are not the same, then the quantiles need not be equidistant. The right column of Figure 23.5 illustrates that the quantiles for two normal distributions with unequal variances will not be equidistant.

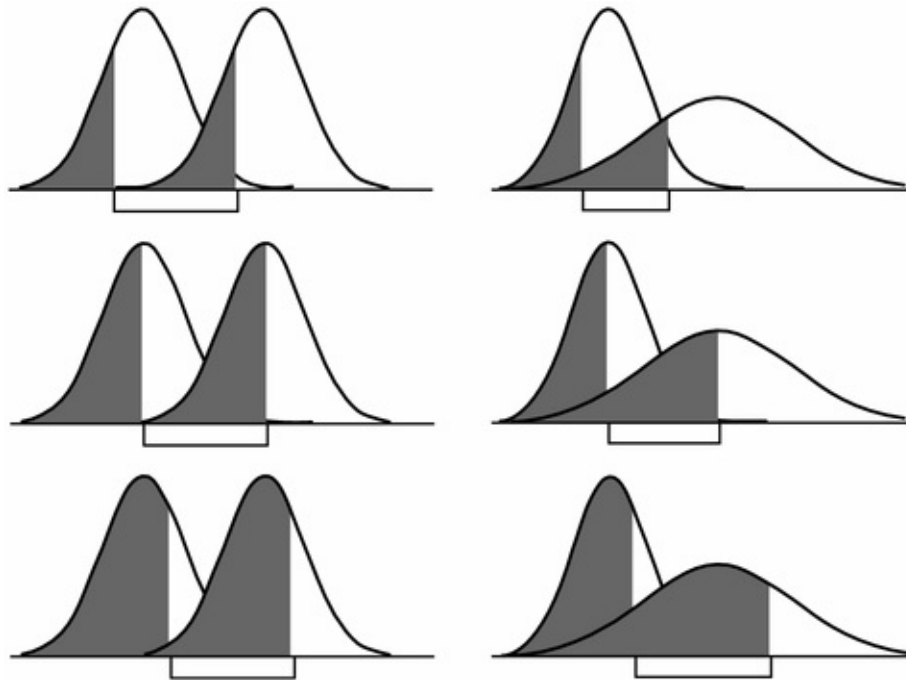


Figure 23.5. Quantiles are equidistant for equal variances (left) and not equidistant for unequal variances (right) (first quartile, median, and third quartile for top to bottom).

Plotting the quantiles of one group against the quantiles of the other group provides a simple method for visualizing whether the quantiles are consistent in their estimates of the mean difference between the groups. Such plots are called quantile-quantile (q-q) plots (the “normal” is dropped because there is no need to assume normality). Figure 23.6 shows the q-q plots for control vs. treatment comparisons for (a) equal variance normal distributions such as those in the left column of Figure 23.5 and (b) unequal variance normal distributions such as those in the right column of Figure 23.5. If the observations were from the same normal distribution (i.e., if the means of the two groups were equal), then the points would fall on the diagonal line with slope 1 and intercept 0. In the left panel of Figure 23.6, the points fall along a line with slope 1 and intercept equal to the mean difference. The q-q points falling on the line indicate that the corresponding quantile difference is consistent with the mean difference estimated from the other quantiles. In contrast, the q-q points in the right panel of Figure 23.6 clearly do not fall on a line with slope 1, indicating that different quantiles provide inconsistent estimates of the mean difference between the two groups. Lower quantiles estimate a negative difference (control exceeds treatment), whereas higher quantiles estimate a positive difference greater than the mean difference (the mean difference determines the intercept of the line).

Note that if the number of observations is unequal in the two groups, then the empirical quantiles of the smaller group are plotted against the corresponding interpolated quantiles of the other group.

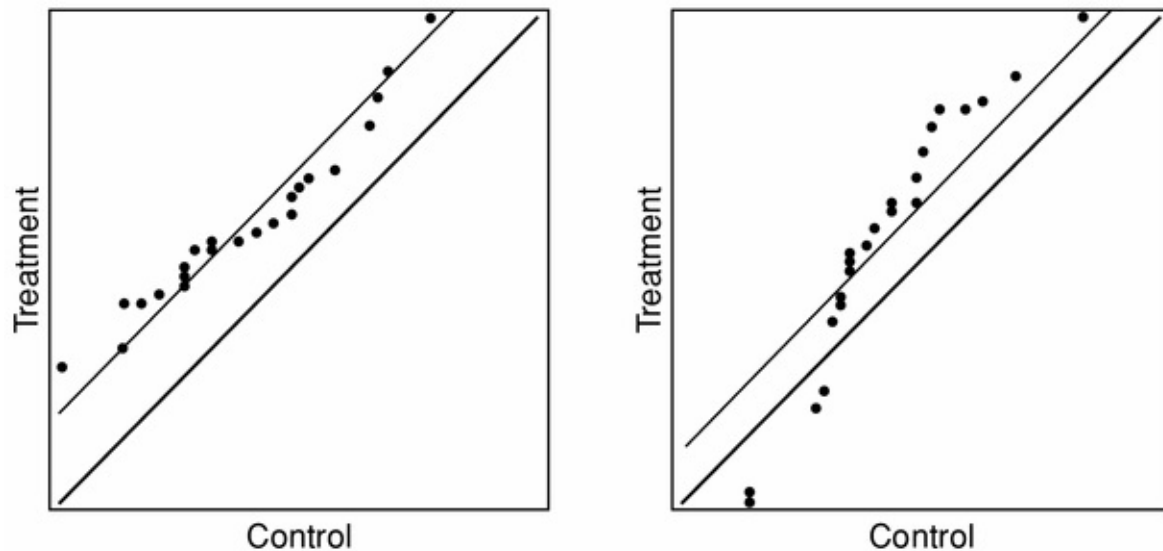


Figure 23.6. Quantile-quantile plot comparing control vs. treatment groups: equal variances (left) and unequal variances (right).

If many of the q-q points are far away from a common line defined by the mean and the standard deviation, then not all the quantiles are telling the same story about whether the treatment mean is higher than the control mean. In other words, q-q points not falling on this common line indicate that the two distributions are not the same. In this case, the two distributions are not the same because they have different variances. Substantially different distributions, whatever the reason, imply that the usual statistical comparison based on Student's *t*-test is inappropriate.

It is a popular myth among psychologists that unequal variances are only a serious problem in the comparison of two groups when the numbers of observations are very different in the two groups. Although it is true that the Type I error will largely be unaffected by the unequal variances (10,000 simulations of the unequal variance comparison of [Figure 23.5](#) yielded a Type I error rate of .051 using a nominal $\alpha = .05$), the Type II error rate and other statistics such as confidence intervals and effect sizes can be adversely affected by the unequal variances. For example, the power for detecting the difference in [Figure 23.5](#) is .97 (based on 10,000 simulations), but falls to .75 for the unequal variances depicted.

No matter what the aggregate statistical tests report, the scientist should strive to get the story right. Figure 23.6 demonstrates that the true story is considerably more complicated than the simple claim that the treatment group, on average, scores higher than the control group does. In this illustration of unequal variances, there may be important consequences of not telling the whole story. Suppose that the treatment were a reading intervention program in the schools. If the data are as in the right graph of Figure 23.6, then the intervention is necessarily deleterious for some, even though it is helpful on average. Although a within-subjects study would be required to know for sure, it is likely for a skill like reading that someone who scores low in the control group would also be among those scoring low in the treatment group. If the variances are unequal as illustrated in Figure 23.6, then the intervention program would be harmful to those most in need of help. To avoid such errors, researchers must tell the whole story represented in their data.

A second method for detecting unequal variances is the *spread-location* or *s-l* plot suggested by Cleveland (1993). Such plots are particularly good for detecting *monotone spread* – a steady increase (or decrease) in the spread or variance of the observations as the location or typical value increases. An *s-l* plot is formed by graphing the absolute errors against the medians of each group. Cleveland (1993) recommends plotting absolute errors and medians because outliers can distort means and variances and therefore interfere with detection of unequal variances. Further, absolute errors are often severely skewed; plotting the square root of the absolute errors removes much of that skew. Figure 23.7 displays *s-l* plots for the same examples used in Figure 23.6. A line connects the two medians of the square roots of the absolute deviations and the positions along the horizontal axis have been “jittered” (a small amount of random error added) so that the points do not overlap. If the variances in the two (or more) groups are equal, then the line connecting the medians of the square roots of the absolute errors will be flat (as in the left panel), but if the variances are unequal, the line will be sloped (as in the right panel). Displaying all the errors provides a visual context for judging the magnitude of the slope relative to the range of errors.

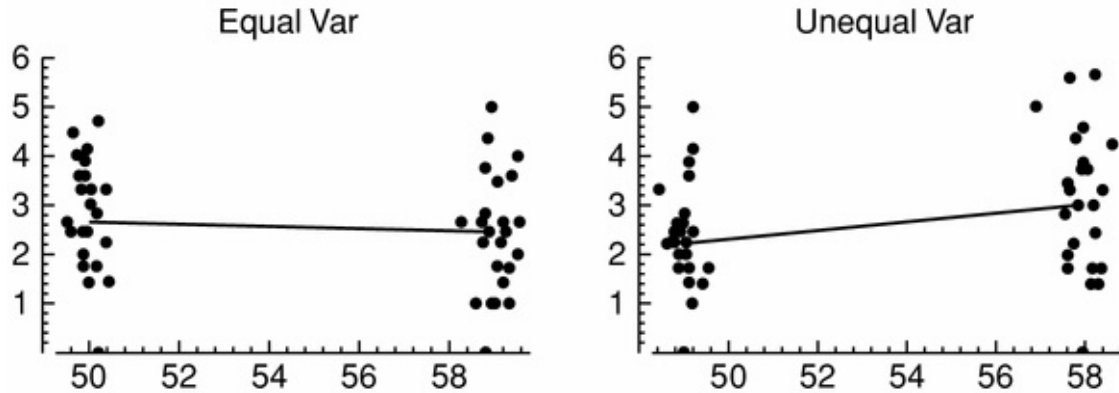


Figure 23.7. Spread-location plots for the examples of [Figure 23.6](#).

Detection of Outliers.

The studentized deleted residual is easily generalized to two (or more groups). The studentized residual is given by

$$r_{ij}^* = \frac{Y_{ij} - \bar{Y}_{[i]j}}{s_{[i]}} \quad (23.5)$$

where j denotes group; this statistic can be compared to Student's t -distribution with $n - 3$ degrees of freedom. A shortcut formula exists that does not require recomputation after each deletion; however, it is easier simply to use a standard regression program (regressing Y on a coded predictor – either contrast, effect, or dummy) that produces studentized deleted residuals.

Multiple Groups

The generalization of the detection techniques for two groups to experiments with multiple groups – using one-way, two-way, or higher factorial ANOVA – is straightforward. For testing normality, the normal q-q plot for all the errors is useful, as are normal q-q plots for the data (or errors) within each group. The spread-location plots are easily extended by using each group's median (or mean) to identify its location in the graph, regardless of the one-way or factorial structure of the groups. And the studentized deleted residual has the same definition and can be compared with the Student t -distribution with $n - g - 1$ degrees of freedom, where g is the number of groups. Again, it is much easier to obtain the studentized deleted residuals from a regression program; see Judd, McClelland, & Ryan ([2009](#)) for details on how to use regression programs to

analyze ANOVA models. Only q-q plots comparing groups do not generalize almost exactly. One could make all $(g)(g - 1)/2$ possible q-q plots comparing one group against another. Also, in factorial designs, one can construct the q-q plot comparing one row against another or one column against another. In a 2×2 factorial design, the interaction compares the positive diagonal cells to the negative diagonal cells; the corresponding q-q plot can often be quite informative. Note that in these multigroup comparisons (e.g., a row against a row), the distributions need not be normal (indeed, if the row variable has an effect, these distributions are unlikely to even be unimodal); however, the assumptions underlying the q-q plot (which do not presume normality) should still hold.

Simple Regression

Anyone who answers a lot of statistical questions for psychologists is always amazed at the number of students and colleagues who arrive with their printouts asking for help in interpreting regression results but who have not looked at any graphs of their data and analyses. The most important diagnostic plot for simple regression is a scatterplot with the regression line superimposed. Anscombe (1973) demonstrated that very different data sets could produce the same regression analysis. The statistical results for all four data sets in Figure 23.8 are identical: $Y = 3.0 + .5X$, $r = .82$, $t(9) = 4.24$, $p < .01$. Yet just a casual inspection of the respective scatterplots reveals that these are very different data sets with very different interpretations. Figure 23.8(a) is probably what is expected for a simple regression with those statistics. Figure 23.8(b) clearly has a curvilinear component. Figure 23.8(c) would be a perfect relationship except for a single outlier and in Figure 23.8(d) there would be no relationship except for a single outlier. Although the data in Figure 23.8(b–d) would be most unlikely to occur in a real study, the point is that a wide variety of data sets – only some of which are consistent with the simple regression model – can produce the same routine statistical analysis. To know what story the data are really telling, it is essential to examine the scatterplot for any simple regression analysis.

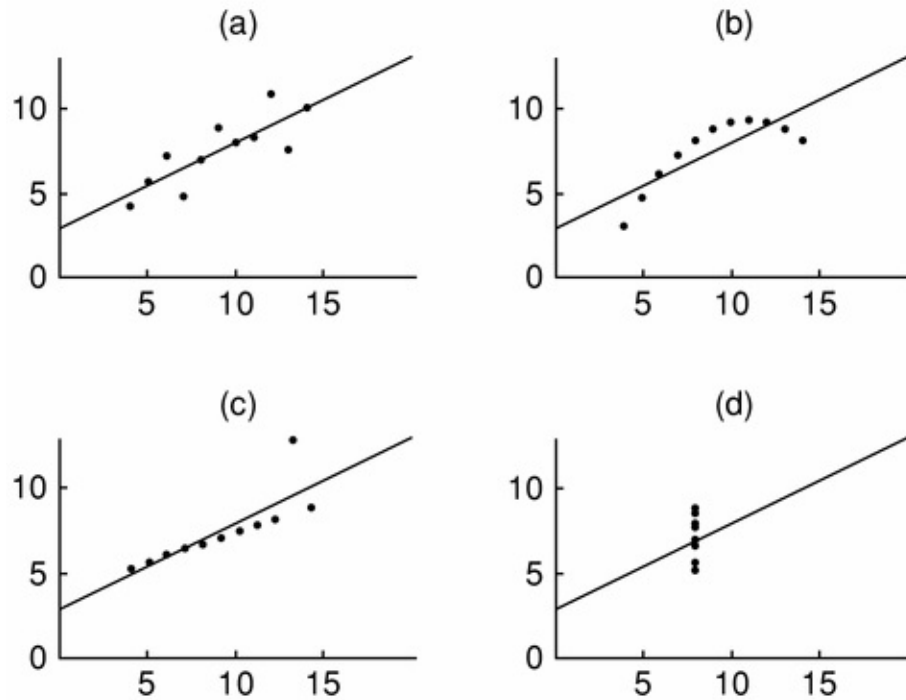


Figure 23.8. Anscombe's (1973) four data sets having the same statistical analysis ($Y = 3 + .5 X$, $r = .82$, $t(9) = 4.24$, $p < .01$).

Outliers.

For simple regression there are three separate outlier questions. Given that there are two variables – a predictor and a criterion variable – two outlier questions are obvious. Is the predictor value for an observation unusual? Is the criterion value unusual? The third question pertains to the joint effect of the predictor and criterion values by asking about the influence of the observation on joint inferences about all the parameters in the model. That is, does the observation distort or have undue influence on the overall regression model? Each outlier issue is considered in turn.

Is the predictor value unusual? All observations contribute equally to estimating the mean; that is, each observation has a weight of $1/n$. In contrast, each observation does not contribute equally to the estimate of the slope in regression. To gain insight into this issue and its importance, it is useful to consider an equivalent, but conceptually different, way to estimate the slope of the best-fitting least-squares line. A common formula for calculating the slope of the regression line is given by

$$(23.6)$$

$$b = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{j=1}^n (X_j - \bar{X})^2}$$

A little algebra yields this equivalent formula for the slope

$$b = \sum_{i=1}^n w_i \left[\frac{Y_i - \bar{Y}}{X_i - \bar{X}} \right] \quad (23.7)$$

$$\text{where } w_i = \frac{(X_i - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}.$$

The term in brackets in [Equation 23.7](#) is the slope “suggested” by each point – the change in Y between the point and the mean for the variable Y divided by the change in X between the point and the mean for the variable X , that is, the slope of a line between the data point and the mean point. The estimated slope for the regression is simply the weighted average of all the individual slopes suggested by each point. [Figure 23.9](#), using the data from [Figure 23.8\(a\)](#), illustrates the individual slopes for each data point.

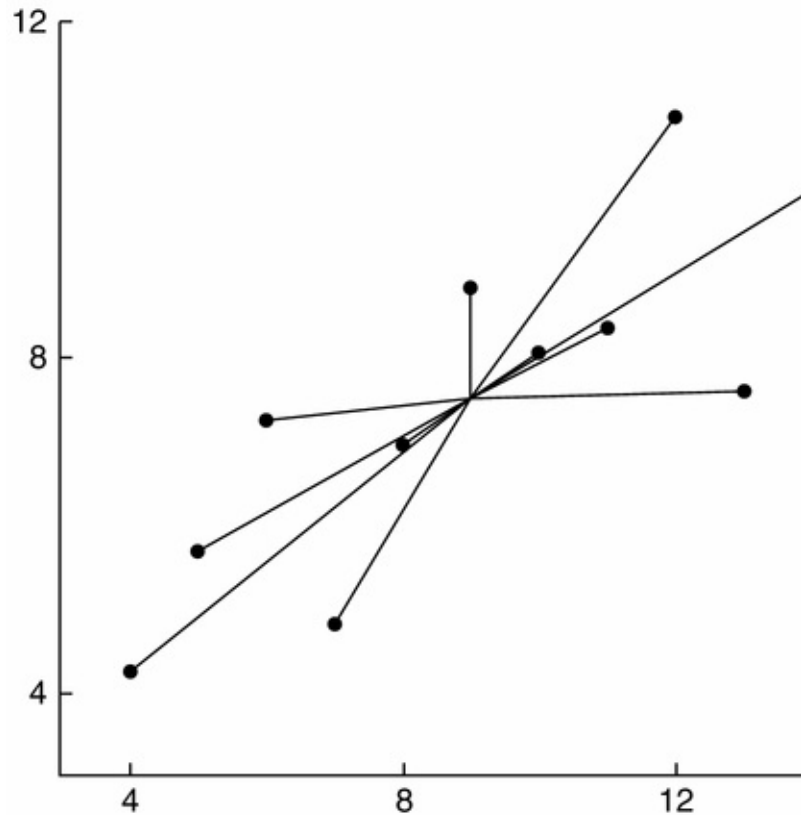


Figure 23.9. Slopes between each data point and the mean point.

The weight given each data point when calculating the overall slope estimate is based on the unusualness of the data point in terms of the predictor variable. The further the predictor variable is from the mean, the greater its weight. This makes sense because the slopes of long lines in Figure 23.9 are unlikely to be affected much by small changes, perhaps caused by error, in the criterion variable Y ; hence, our confidence in the estimated slope is greater for long lines. In contrast, small changes in Y for short lines could dramatically alter the slope; hence, our confidence in the estimated slope is much less for short lines. The details of the calculations are presented in Table 23.2. One can think of each observation as having a vote on the slope that is to apply to all the observations, but with the votes of some observations counting more. If all the observations are telling more or less the same story, then all the observations ought to be voting for essentially the same slope. In the case of a perfect relationship, all the individual slopes would be identical and would equal the overall slope. In Figure 23.9, where the data had a correlation of .82, the individual slopes vary from .02 to infinity. However, those observations with extreme slopes generally receive little weight. In particular, the infinite slope receives a weight of 0. The weight is

a measure of how much impact or leverage an observation has on the overall slope estimate. In this case, four observations (3, 6, 8, and 11) account for approximately 75% of the total weight. It would be undesirable for a single observation to have a very large percentage of the total weight, because in that case the slope votes of all the other data points are ignored in calculating the “overall” slope. In that case, the “overall” slope is really a description of only one data point. This occurs when one observation has a very unusual (relative to the other observations) predictor value; such is the case in [Figure 23.8\(d\)](#).

Table 23.2. Slope Calculations for the Data Set of Figure 23.8(a)

	X1	Y1	slope	Wt	Wi *slope
	10	8.04	0.54	0.01	0.00
	8	6.95	0.55	0.01	0.01
	13	7.58	0.02	0.15	0.00
	9	8.81	inf	0.00	0.00
	11	8.33	0.41	0.04	0.02
	14	9.96	0.49	0.23	0.11
	6	7.24	0.09	0.08	0.01
	4	4.26	0.65	0.23	0.15
	12	10.84	1.11	0.08	0.09
	7	4.82	1.34	0.04	0.05
	5	5.68	0.46	0.15	0.07
Sum	99	82.51		1.00	0.50
Mean	9	7.50			
Sum Sq Dev	110				

Most modern regression programs report the lever (sometimes unhelpfully referred to as the diagonal of the hat matrix) h , which represents the weight or leverage an observation has in determining the overall model. For simple regression, the lever is given by

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}. \quad (23.8)$$

The lever represents an observation's weight in determining the overall model.

In simple regression, the model consists of two parts: the mean (or intercept) and the slope. Each observation has equal weight in the voting for the mean and weight proportional to the squared deviation of the predictor variable from its mean in the voting for the slope. The lever, as shown in [Equation 23.8](#), is simply the sum of those two weights. Unusually high values for the levers are undesirable because it means that what the regression program reports really only applies to the one or two observations with unusually high levers. It is inappropriate and misleading to report a regression equation as if it applies to all n observations if, because a few points have high levers, it actually only applies to a few observations. In judging the magnitude of levers, it is useful to note that the sum of the levers necessarily equals the number of parameters, 2 for the case of simple regression. Hence, the average lever equals $2/n$ for simple regression. If a single lever is near 1, then it implies that one of the two parameters of the simple regression model is allocated to predict that single observation; this is clearly undesirable. An example is the unusual observation in the data set of [Figure 23.8\(d\)](#); its leverage value is 1.0, whereas all the other observations have leverage values of 0.1. That one unusual observation effectively determines the slope.

It is well known that restricting the range (more properly, restricting the variance) of the predictor variable attenuates the correlation. The converse applies when there is a single outlier with an extreme predictor value. That one observation artificially increases the range (variance) and thus inflates the correlation. Allowing one observation to inflate the correlation, or even to create one when otherwise there would be no relationship, obviously increases the chances of making Type I errors – rejecting the null hypothesis when the null hypothesis is in fact correct. Observations with unusual predictor values, assuming they do not also have unusual criterion values, often make regression models appear better than they actually are. In those cases, the story told by the regression model really only pertains to that one observation, and it is very misleading to pretend that the regression story applies to all the data.

Is the criterion value unusual? This is the same outlier question we asked with respect to group comparisons and the answer is the same. The studentized deleted residual is defined as before as

$$\mathbf{r}_i^* = \frac{Y_i - \hat{Y}_{[i]}}{\mathbf{s}_{[i]}} \quad (23.9)$$

Again, there are shortcut formulas, but these are seldom needed because most modern regression programs report, if requested, the studentized deleted residuals. For simple regression, these can be compared to Student's t -distribution with $n - 3$ degrees of freedom. Conceptually it is important to note that the studentized deleted residual is equivalent to the value of the t -statistic one would obtain for testing the coefficient of a dummy variable added to the regression model with a value of 1 for the indicated observation and a value of 0 for all other observations. In other words, the studentized deleted residual asks whether the observation is so unusual, relative to the others, that it is worthwhile to add a separate parameter to the regression model just to account for that one observation. As an example, the studentized deleted residual for the unusual observation in the Anscombe (1973) data set of Figure 23.8(c) equals 3.9, while the next highest (in absolute value) studentized deleted residual is only -0.9 .

A single unusual criterion value can dramatically inflate the variance of the criterion, variance that the predictor variable cannot possibly explain. This has the effect of greatly reducing the correlation and associated test statistics. A single unusual criterion value therefore greatly increases the chances of a Type II error – not rejecting the null hypothesis when the null hypothesis is in fact false. Observations with unusual criterion values, assuming they do not also have unusual predictor values, often make regression models appear much worse than they actually are and greatly reduce statistical power.

Does the observation have undue influence on the overall regression model? An observation might have an unusual predictor value, but if its criterion value falls near the regression line determined by the other observations, then it will not unduly influence the estimates of the slope and intercept of the regression line. Similarly, although it will increase the mean squared error substantially, an observation with an unusual criterion value will have little effect on the estimates of the slope and intercept if it has a typical predictor value. The really nasty observations – the ones that have greatly disproportionate influence on the overall model – are those that have both predictor and criterion values that are at least a little bit weird. It is therefore useful to have an index that is the product of a function of leverage and a function of the residual, standardized or studentized. One popular such index is Cook's (1979) D , which is defined as

$$D_i = \frac{r_i'^2 h_i}{2s(1 - h_i)} \quad (23.10)$$

Many computer programs will produce Cook's D , if asked, so this formula is more important for its conceptual than computational value. To be large, Cook's D requires that both the squared standardized residual and the lever h be reasonably large. Large values are those that stand out relative to the other values; a large gap between the largest Cook's D and the next-largest value usually indicates a serious problem. As illustrated in an example later, an observation or two with large values of Cook's D can seriously distort the parameter estimates; if so, the effects on Type I and Type II errors are unpredictable. In short, observations with undue influence are very nasty and can really ruin an analysis. Watch out for them!

Many computer programs now provide an overwhelming choice of outlier indices. In general, the indices can be sorted into categories corresponding to one of the three outlier questions. The subtle differences among indices for the same question are generally not important; be sure to use one index from each category.

Homogeneity of Variance.

Many regression programs now provide a plot of the residuals against the fitted or predicted values from the regression equation. The most common violation of homogeneity of variance observable in these plots is a funnel shape – little spread for small predicted values and much greater spread for large predicted values. However, it is often easier to see the changing spread of the residuals as a function of the predicted values if the residuals are “folded” by either squaring them or taking the absolute values. Just as for the two-group comparison presented earlier, many find that spread is easiest to detect visually if the square root of the absolute value of the residuals is plotted against either the actual or predicted data values. One can even fit a regression line (or a robust curve) to verify any visually apparent change in the spread.

Figure 23.10 presents examples of the plots for detecting violations of homogeneity of variance. The plots in the left column of Figure 23.10 pertain to data generated using a fixed error distribution for all observations, while those in the right column pertain to data generated using an error distribution proportional to the size of the respective predicted values. In the usual regression scatterplot (the top row of Figure 23.10), the data values should be scattered evenly about the regression line. It is possible, but somewhat difficult, to observe that the scatter about the regression line is increasing in the top-right plot. Plots of the errors or raw residuals against the predicted values are in the second row

of Figure 23.10; the errors in these plots should have a “cloud” appearance. However, the graph in the right column has an obvious “funnel” shape, indicating a clear violation of the homogeneity of regression assumption. Finally, spread-location plots are in the bottom row of Figure 23.10; these “folded” errors [$\sqrt{|\text{error}|}$] should have no apparent trend as the predictions increase. That is the case in the left column where the dotted regression line essentially has a slope of zero, but is definitely not the case in the right column where the folded errors steadily increase with the size of the predictions. Such graphs provide an easy visual method for detecting violations of the homogeneity of variance assumption in regression.

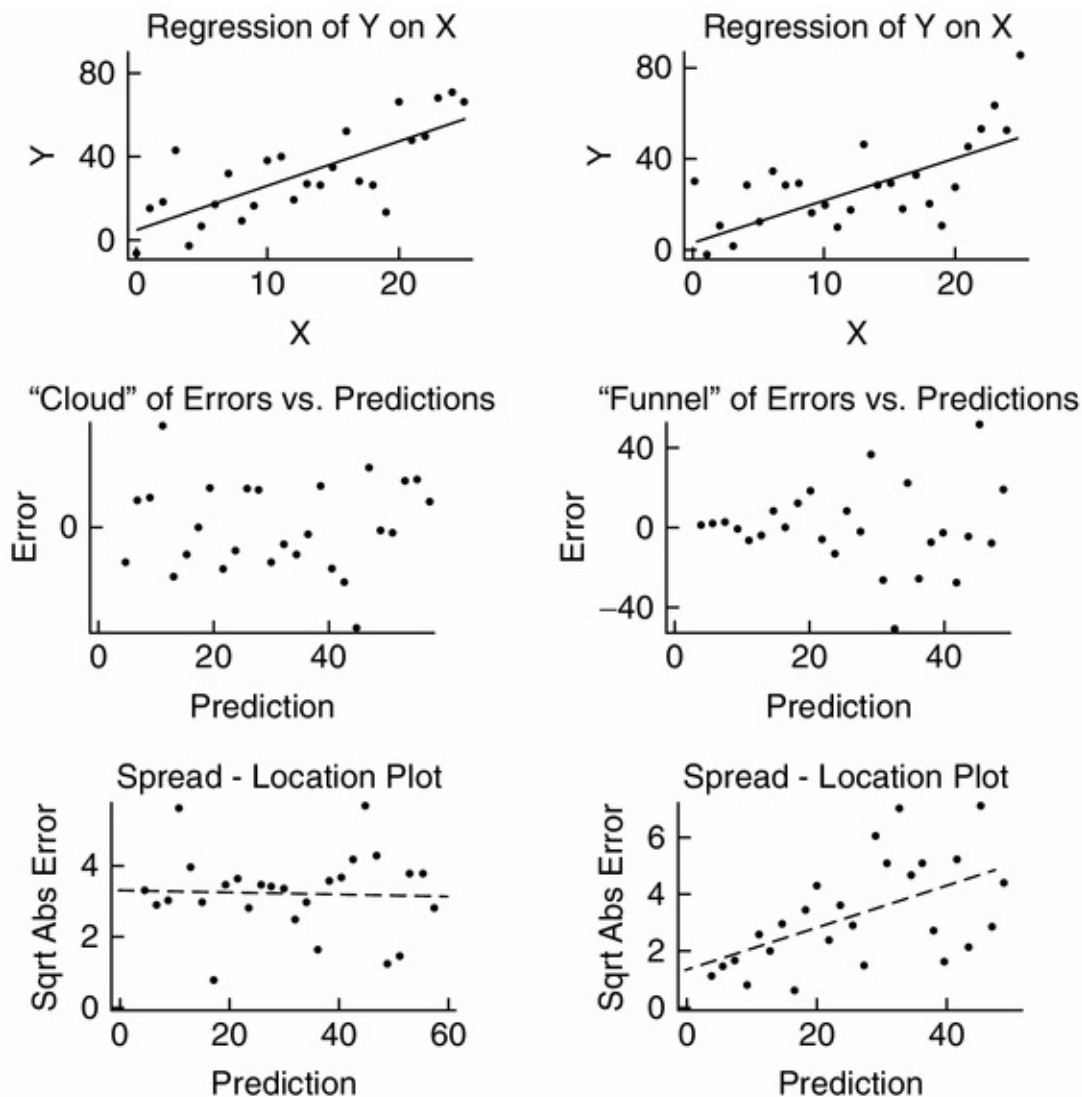


Figure 23.10. Plots for detecting violations of homogeneity of variance in regression analysis.

Multiple Regression

The techniques for detecting nasty observations for simple regression generalize easily to multiple regression. The generalizations of the formulas for levers, studentized deleted residuals, and Cook's D are not practical for hand computation, so we do not present them. Most modern regression programs will produce those outlier indices, or equivalent ones that answer the three outlier questions for regression. The interpretations of the studentized deleted residuals and Cook's D are exactly the same as for simple regression. There is one additional wrinkle for leverage in multiple regression. We could assess whether each predictor value is unusual with respect to that predictor. However, doing so we might miss an observation whose predictor values were not unusual within each predictor but whose pattern across all the predictors was unusual. For example, if we were using actual height and weight to predict satisfaction with body image among a sample of adolescent girls, a height of 5' 9" would not be particularly unusual and a weight of 95 pounds would not be particularly unusual, but the combination would be. Thus, we would like one index to assess the unusualness of the whole pattern of an observation's predictor values. The generalization of the leverage index h detects this kind of unusualness quite well.

The scatterplot of simple regression obviously does not generalize well for multiple regression. However, the partial regression leverage plot is a type of scatterplot for each variable that is useful in identifying unusual observations and relationships in multiple regression analyses. For a particular predictor X , instead of plotting Y against X , the partial regression leverage plot graphs the residual Y against the residual X after both Y and X have been predicted by all the other variables in the multiple regression model. The best fitting slope in this plot equals the regression coefficient for that variable in the full multiple regression model. Model deficiencies such as curvilinearity and unusual data points are often easy to spot in such plots (Velleman & Welsch, 1981 provided good examples). Many regression programs have options for generating the complete set of partial regression leverage plots.

Remedies

Once nasty and unruly data have been detected using the diagnostic graphs and statistical indices described earlier, it is important to take remedial action to prevent them from ruining the analysis. In this section we consider several remedial strategies that are likely to be effective. However, we first consider a

commonly tried strategy – nonparametric statistics – that is not likely to be either effective or appropriate. It is ironic that an ineffective remedy like nonparametric statistics is accepted without question by social psychologists while effective remedies such as transformations and removing outliers remain controversial.

Nonparametric Statistics

When confronted with unruly, nasty data, many researchers turn to nonparametric statistics in the belief that all the strong assumptions required by statistical tests based on least-squares are relaxed. In fact, the only assumption that is relaxed is the normality assumption; assumptions about having the same distribution and homogeneity of variance still apply. For example, the Mann-Whitney U (also known as Wilcoxon Rank-Sum) and the Kruskal-Wallis tests – the nonparametric analogs to two-sample t -tests and one-way ANOVA – allow the data to have a weird distribution, but the same weird distribution must apply to the data within each group. When the data are nasty because of a few outliers, it is unlikely that each group will have the same nonnormal distribution. Most statistics textbooks written by psychologists fail either to describe the necessary assumptions underlying nonparametric tests or to warn of their reduced statistical power. As a consequence, nonparametric statistics are often used inappropriately in the psychological literature.

Note that the derivation of the q-q plots in [Figure 23.5](#) still applies if the normal distribution is replaced by any other distribution. If two groups have distributions of the same shape, shifted only in location, then equivalent quantiles from each distribution should estimate the location shift. Thus, even with nonnormal distributions, plots of one group's quantiles against the other group's should yield a straight line parallel to the diagonal. The spacing of the points along the line would not be the same as if the data were from a normal distribution (e.g., the spacing between quantiles in the tail of a skewed distribution will be further apart), but the q-q line will be straight and parallel to the diagonal. Thus, any problems detected by examining q-q plots cannot possibly be remedied by using nonparametric statistics, because the assumptions of such tests are necessarily violated.

Transformations

Nonlinear but monotonic transformations of the criterion or dependent variable can often correct violations of normality and homogeneity of variance. As an added bonus – Emerson ([1991](#)) and Tukey, Mosteller, and Hoaglin ([1991](#)) even

argue it is the more important benefit – data transformed to solve such problems often yield a simpler model (i.e., fewer interactions and polynomial components). There are types of data that one can anticipate will benefit from transformation. Data that are constrained at one or both ends of their ranges are likely to violate the normality and homogeneity assumptions. Examples of such data include counts, completion times, proportions, and correlations (when used as data themselves). Generally the constrained end of the scale needs to be stretched and/or the unconstrained (or skewed) end needs to be pulled in. Appropriate transformations in these cases include square root for counts, logarithm for times, arcsin or logit for proportions, and Fisher's Z for correlations (see Judd, McClelland, & Ryan, 2009 for more details about these known transformation problems and Bargh & Chartrand, Chapter 13 in this volume, for reaction time transformations). Data spanning several orders of magnitude are also likely to be problematic for the standard analysis assumptions. For example, when studying vocabulary size of preschool children, the difference between knowing, say, 50 and 100 words is surely not the same as the difference between knowing 1,050 and 1,100 words. However, using untransformed word counts presumes such differences are equivalent.

Box and Cox (1964) and Tukey (1977) have popularized the family of power transformations, which are frequently useful for solving problems of nonnormality and heterogeneity of variance. Instead of analyzing the dependent variable Y , one analyzes

$$Y(\lambda) = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & (\lambda \neq 0) \\ \log Y & (\lambda = 0), \end{cases} \quad (23.11)$$

for an appropriate value of λ . The power transformation is undefined for the power zero, but conveniently the log function fills that role because it is the limit of the power function as the power approaches zero. Values of the power transformation are displayed in Figure 23.11 in what Tukey (1977) refers to as the “ladder of powers .” The subtraction of 1 and division by λ serves the purpose of making the power transformation curves be aligned in Figure 23.11 at $Y = 1$; in practice, the extra arithmetic is not needed. Values of λ smaller than 1 pull in the tails of positively skewed distributions, while powers greater than 1 pull in the tails of negatively skewed distributions. There are sophisticated statistical and graphical procedures for estimating the optimal power

transformation for achieving normality and homogeneity of variance (see Madansky, 1988 for a review of many such procedures), but simply trying a few powers usually suffices to determine whether a transformation will help. That is, one usually generates a few alternative power-transformed criterion variables using, say, the whole number powers between -2 and $+2$. The diagnostic plots and indices described in this chapter are then generated for an analysis of each power-transformed criterion variables. The approximately best power – the one that produces an analysis with the fewest assumption violations and outliers – is generally clear. One could then repeat the process with a finer range of powers around the approximate best, but such precision is usually not necessary.

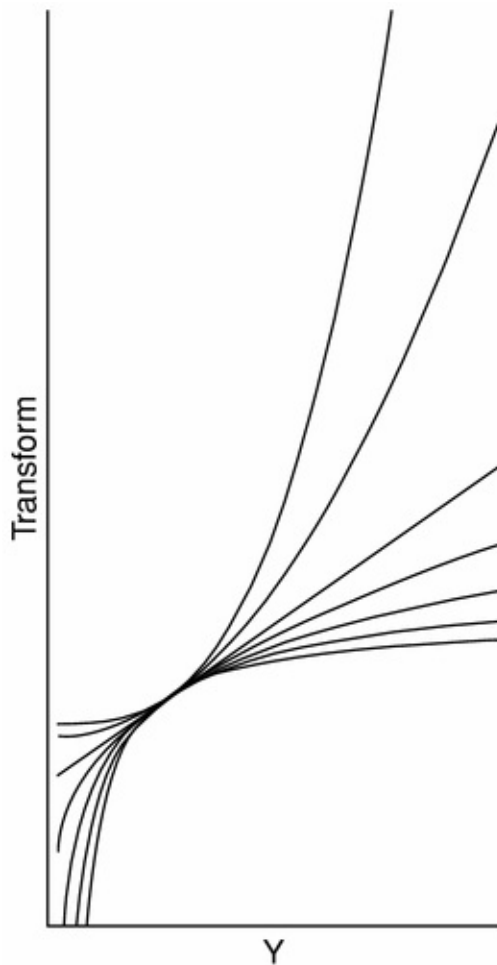


Figure 23.11. Family of power transformations. (Powers, descending on the right edge of the figure, are 3, 2, .5, 0, $-.5$, and -1).

Transformations are sometimes controversial in social psychology, but they should not be. Statisticians readily accept the need for transformations (e.g., Atkinson, 1985). There is seldom reason to believe that the scale we happen to

use ought to be linearly related to what we are trying to measure. In many areas of psychology where enough individual data have been obtained to assess linearity, nonlinear functions are the rule. Examples include psychophysical functions in perception and utility functions in decision making. There is no reason to presume that social psychology will escape such nonlinearity in our measurement scales. Psychologists sometimes mistakenly assume that because their response scale is measured in physical units that are interval or ratio scales they need not worry about linearity. However, just because, for example, times to solve problems are measured on a ratio scale of time does not imply that those times represent even a linear scale of a psychological concept such as problem difficulty (see Judd & McClelland, 1998 for further discussion of this issue.)

Psychologists who think they are opposed to transformations are often surprised to learn that nonparametric statistics, of which they approve, are actually a transformation approach to data analysis. For example, Conover and Iman (1981) demonstrated that common nonparametric statistics are essentially equivalent to applying the usual least-squares methods to rank-transformed data – a transformation of the data that is usually less gentle than the family of power transformations. In this sense, power transformations offer an intermediate strategy between analyzing the raw data with problems of nonnormality and heterogeneity of variance and using nonparametric statistics. Anyone willing to use nonparametric statistics ought to be willing to consider the gentler power transformations. Transformations can often improve the health of sick data, and social psychologists ought to consider them more often in data analysis.

Sometimes it is impossible to find a transformation that both achieves normality and stabilizes the variance. Judd, McClelland, and Culhane (1995) provide an overview of some more advanced transformation and analysis techniques to deal with such situations. However, if the data are so messy that a simple power transformation does not improve things, then most social psychologists ought to seek help from an expert because advice from a primer like this one will not be sufficient.

Outliers

Unfortunately, dealing with outliers is sometimes even more controversial than using transformations. If the identified outlier turns out to be the result of a typographical error or an impossible value (e.g., a meter reading in an energy conservation study that implies negative consumption since the previous reading), then almost no researchers would object to discarding or, if possible,

correcting the data value. The controversies arise when the outliers have no obvious procedural explanation. In the early days of psychological research when identification of outliers was usually based on ad hoc rules proposed by the investigator, researchers were justified in being skeptical of such fiddling with the data. However, we now have principled, statistically based methods for identifying outliers. The outlier detection methods are no more and no less sound than the methods, say, for determining if there is a difference between two means. While it certainly is unethical to delete observations just to get the result one wants, it is equally unethical to present a model (or story) for the data that really is determined by just a few of the observations (e.g., reporting a linear regression for the data in [Figure 23.8\[d\]](#)).

As of this writing, there is growing concern in social psychology about what has been labeled as “questionable research practices” (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). The concern is that these questionable research practices encourage massaging the data and collecting more data and doing many little studies until achieving a significance level to satisfy journal editors and reviewers. One such practice is the *arbitrary* removal of outliers until significance is achieved. For example, using an arbitrary cutoff such as data values more extreme than two standard deviations. However, the methods in this chapter are not arbitrary and are instead based on sound statistical principles. Identifying a significant outlier using the studentized deleted residual is no more or less arbitrary than any significance test. The concern about questionable research practices is focused on how the removal of outliers might produce spurious significance. As shown in this chapter, outliers with extreme levers or large values of Cooks’ D can produce significant results when otherwise there would be none. Ignoring these kinds of outliers should be added to the list of questionable research practices because doing so can produce isolated significant results that are impossible to replicate. Doing a principled outlier analysis as detailed in this chapter is *not* a questionable research practice. To the contrary, examining outliers and assumptions should be a part of any quality data analysis. Of course, the researcher should be totally transparent about how the outlier analyses were conducted and how they were remedied.

Dealing with outliers may be less controversial if we think in terms of a story metaphor. The model resulting from data analysis tells a story about our data. All observations ought to help us tell that story. It might be a very complicated story with subplots (interactions). If an observation is part of that story then it should not be necessary to add a dummy parameter to the model just for that observation. If it is necessary to add such a parameter (i.e., if the studentized

deleted residual is large), then that observation is telling its own story (the dummy parameter is needed for that part of the story) and it is a different story than the other observations are telling. It is wrong to pretend that this observation is part of the same story as the other data. Even worse, if that observation has high leverage and a large Cook's D , then it will substantially distort the story told by all the other observations. The result may well be a story that does not apply to any of the data. At the very least, researchers should examine their data for outliers and then report analyses with and without the outliers included. Otherwise there will be doubt as to what story the data are really telling.

Researchers in social psychology are often justifiably pleased when they are able to account for significant but small proportions of the overall variance in their data. So too we should be pleased if we are able to provide models that provide good accounts of the data for a large proportion of our observations. If we are not embarrassed when we say that we understand, say, 15% of the variance in our data but that we do not have a clue about the other 85%, then we certainly should not be embarrassed to admit we have a model of the data that applies to, say, 90% of our observations but we do not have a clue what the other 10% of the folks (outliers) were doing. A good model for most of our data is better than a poor model for all of our data.

Finally, it is important to identify those observations telling different stories, not so that we can discard them but so that we can listen to the different stories they have to tell. For example, to design intervention programs it would be useful to find resilient children who had thrived and excelled despite being raised in adverse environments. Identifying outliers and figuring out why they are outliers often leads to new theoretical advances. That cannot occur if the outliers are left submerged within the story told by the other observations. For all these reasons, social psychologists should always examine their data for outliers.

Example

An example will help elucidate the diagnostic and remedial tools described in the preceding section. These data pertain to the relationship between the number of grants in the physical and social sciences awarded by the National Science Foundation (NSF) to various universities. These data were assembled by a colleague whose university's administration was concerned about the low number of NSF grants they received. They had decided to remedy the situation

by providing extra resources to departments in the physical sciences. This colleague hoped to convince his administration that it was equally important to support social science departments because there was a strong relationship between the number of grants received in the physical and social sciences at peer institutions. The data analyzed here are the number of NSF grants of each type for the given state's two primary public universities and their self-identified peer institutions. Each university is categorized as either the state's flagship or land grant university. Figure 23.12 shows the data and best-fitting regression lines for the flagship and land grant universities. Despite the apparent divergence of the slopes, the difference in slopes (i.e., a test of the interaction between university category and number of social science grants) is not significant, $F(1, 15) = 1.9$, $PRE = .11$, $p = .19$ (note that PRE is the proportional reduction in error variance, equivalent to the squared partial correlation, when adding the tested predictor to a model containing all the other predictors; see Judd, McClelland, & Ryan, 2009). However, the plot reveals a potential outlier – a flagship university that has by far the greatest number of physical science grants – that may be distorting the regression line for the flagship universities. In any case, a visual check of the regression assumptions as well as an outlier analysis is needed.

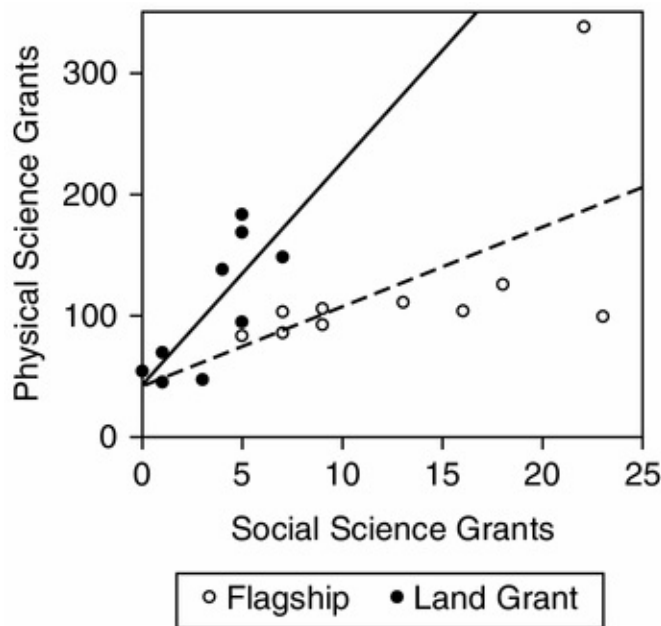


Figure 23.12. Relationship between NSF physical science and social science grants at flagship and land grant peer institutions.

Figure 23.13 displays the normal quantile plot (on the left) and the spread-location plot (on the right). One point, which corresponds to the apparent outlier

in Figure 23.12, is far away from the normality line in the normal quantile plot; its steep slope relative to the other points identifies this as a normality violation that could have substantial impact on the analysis. The spread-location plot clearly shows that the square root of the absolute value of the residuals is increasing with the size of the predicted values from the regression; thus, the assumption of homogeneity of variance is violated. The studentized deleted residual for the unusual observation is 6.78, its leverage is .32, and its Cook's D is 1.33, about twice as large as the next value of Cook's D . Its lever does not identify it as a particularly unusual observation in terms of its predictor values. However, the large studentized deleted residual ($p < .0001$) suggests that this university is telling a very different story about the relationship between the number of physical and social science NSF grants; the large value of Cook's D indicates that it is having a disproportionate effect on the overall model. If the outlier is omitted, the interaction is significant, $F(1, 14) = 15.38$, $PRE = .53$, $p = .0015$. In other words, the story we tell about whether the relationship is different for flagship and land grant universities depends entirely on whether we include this one university. We should not allow one observation to dominate the story we tell about all the data! In this particular case, the unusual university has by far the highest total number of grants. And its number of physical science and social science grants fits neither the pattern of the land grant universities nor the pattern of the other flagship universities. In retrospect, this might not be a legitimate peer institution; it may have been included, wishfully, as a peer only because it was in the same intercollegiate athletic conference. For all these reasons, it is appropriate to conduct an analysis with that outlier removed. If the analysis changes appreciably without that university, that of course does not prove any of our post hoc suppositions made earlier. Instead, those suppositions might provide hypotheses to be explored in a larger study including all the state universities.

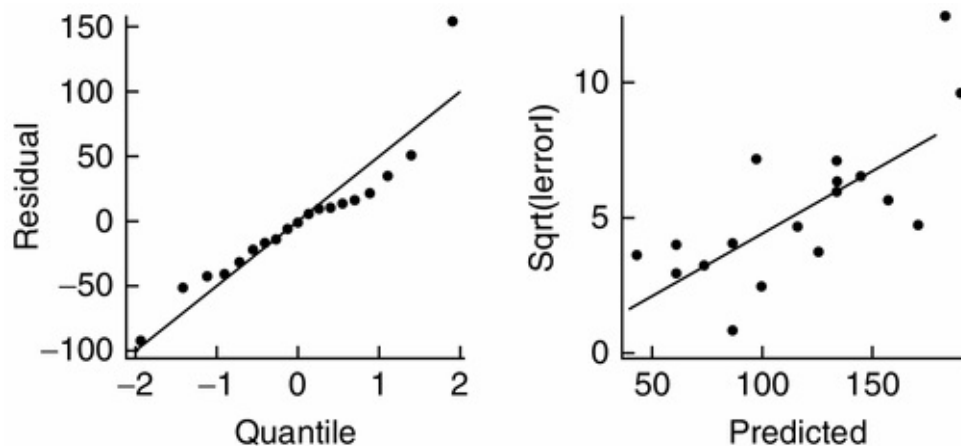


Figure 23.13. Normal quantile and spread-location plots for the untransformed analysis of the NSF grant data.

Removing the one clear outlier does not repair the violation of homogeneity of variance apparent in the right panel of [Figure 23.13](#) (the new graph is not presented here). Hence, a transformation may be appropriate. Also, there are a priori reasons for anticipating the need for a transformation of these data. The scale for the number of grants is not likely to be linear with an underlying scale of institution quality. For example, is the functional difference between, say, 5 and 10 grants equivalent to the difference between, say, 105 and 110 grants? We are likely to judge the second difference to be negligible while considering the first difference to be quite large. Analyzing the raw data implicitly treats these two differences as if they were equal. Also, counts are likely to follow a Poisson distribution rather than a normal distribution. In a Poisson distribution the variance is a function of the mean level. The square root transformation is well known for removing this dependency for count data. Although one need not transform on both sides, it seems appropriate in this case to transform both the criterion (number of physical science grants) and the predictor (number of social science grants). [Figure 23.14](#) shows the normal quantile plot of the residuals and the spread-location plot after the square root transformation has been applied with the outlier omitted (it remains an outlier, although not as extreme, after the transformation). Both plots suggest that the analysis of the transformed data without the outlier reasonably satisfy the normality and homogeneity of variance assumptions. Any weaker power transformation (i.e., a power between 0.5 and 1) leaves a positive slope in the spread-location plot, whereas any stronger transformation (such as the log or the inverse) induces a negative slope. Hence, the square root transformation (power = 0.5) is best for correcting the variance problems in these data.

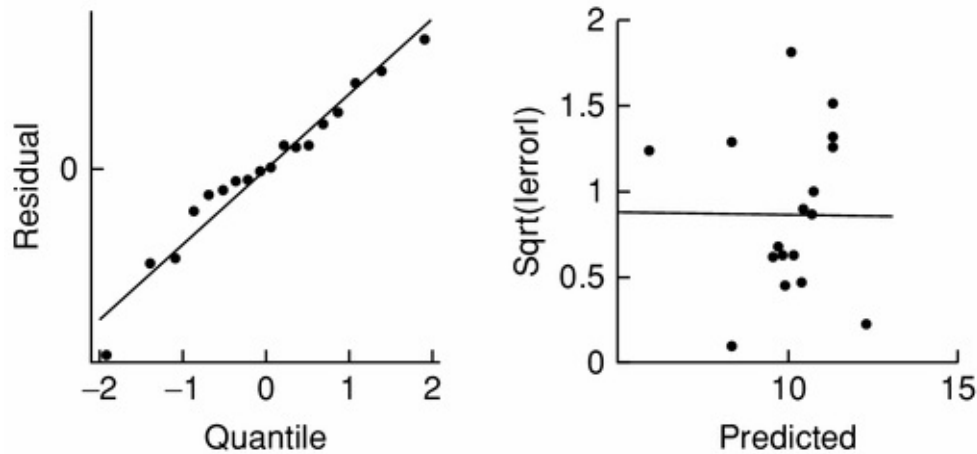


Figure 23.14. Normal quantile and spread-location plots for the square root transformed analysis of the NSF grant data with the outlier omitted.

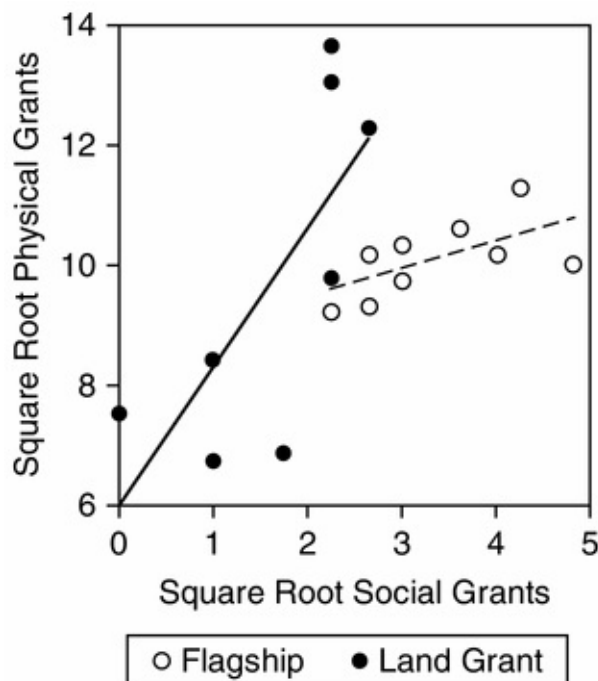


Figure 23.15. Final analysis of square root transformed data with one outlier omitted.

An analysis of the transformed data reveals a second problematical observation – a land grant university that has 56 NSF grants in the physical sciences but none in the social sciences. Some land grant universities have remained closer to their roots as agricultural and mechanical universities and so do not have the full complement of social science departments; this university may be such an instance. Fortunately, the story does not change appreciably if

this observation is included or omitted. So, we will stop with the analysis of the square root transformed data with the first outlier omitted; this final analysis is depicted in [Figure 23.15](#). There is no evidence for a relationship between the number of physical and social science grants at flagship universities (slope = 0.46), $F(1, 14) = 0.6$, $PRE = .04$, $p = .45$, but there is a relationship for land grant universities (slope = 2.4), $F(1, 14) = 15.65$, $PRE = .53$, $p = .0014$. The difference between the two slopes is statistically significant (slope difference = 1.93), $F(1, 14) = 5.15$, $PRE = .27$, $p = .04$. (Omitting the second problematical observation would strengthen the land grant relationship and enhance the slope difference.) Transforming the original data and omitting the outlier yielded not only an analysis that satisfied the important statistical assumptions but also a clear and consistent story for these data. Including the outlier and not transforming the data yielded an analysis that violated the major assumptions underlying the analysis and produced a muddled and inconsistent story for the data.

Conclusion

Many social psychologists, including reviewers of an early draft of this chapter, when first considering remedies for outliers and assumption violations ask: “Isn't there a danger these methods can be abused?” Absolutely, any statistical method can be and probably has been abused by unscrupulous scientists. But that does not imply these methods for detecting outliers or assumption violations and remediating them should be denied to those careful social psychologists who want to understand everything their data have to say. Furthermore, the concern about abuse should not be one-sided. It is a serious abuse to report classical statistical tests when the data contain serious outliers or when the assumptions are substantially violated. In short, scientists seriously mislead their readers when they pretend that one consistent story applies to their data when the methods presented in this chapter would reveal multiple stories inconsistent with the main story. Unwittingly, many of us are guilty of that kind of abuse. The one-sided concern about possible abuses of outlier detection is akin to only worrying about Type I errors. Just as we are frequently reminded to also be concerned about Type II errors, so too we should be reminded that ignoring outliers and assumption violations is also a serious matter.

Probably the greatest psychological obstacle to using outlier detection methods is a belief that it is unethical to alter data for any reason. Although modern outlier detection statistics do provide a principled, rather than arbitrary,

method for identifying observations that do not belong in an analysis, that is not the most important reason for their use. In programmatic research it is those unusual cases telling different stories that lead us to the next insights or that help us to sharpen our theories. If we do not use outlier methods to highlight those unusual cases, then we cannot listen to the important stories they are trying to tell us. The ultimate worth of remedies such as removing outliers or transforming variables is not whether they solve a nasty data analysis problem in a single study, but rather whether they lead to a better understanding or a better theoretical account of an integrated series of studies. As Bacon told us long ago, mature theories facilitate the detection of deviations and in turn those deviations help us improve our theories. Rather than closing our eyes for fear of imagined abuses, we should instead gladly open our eyes wide to look for outliers, nonnormality, and heterogeneity of variance. Doing so will inevitably improve our understanding of our data.

In summary, nasty, unruly observations can ruin one's analysis. A few outliers can grab the analysis so that the resulting story applies to none or only a few observations. The typical statistical output from computer programs usually provides no clues that the analysis has been overwhelmed by a few observations or by systematic violations of the assumptions of normality and homogeneity of variance. However, if prodded, most modern data analysis programs will provide useful diagnostic plots and outlier indices. There is no excuse for not using these diagnostic tools. Failure to detect important assumption violations and outliers may mean that researchers report misleading stories about their data. Once nasty, unruly data are detected, they can often be tamed through remedies such as transformations and the deletion of outliers. Principled methods exist for identifying appropriate transformations and marking outliers for possible deletion. Using those methods will mean that more data stories have happy endings.

References

- Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*, 68, 13–20.
- Atkinson, A. C. (1982). Regression diagnostics, transformations and constructed variables (with discussion). *Journal of the Royal Statistical Society B*, 44, 1–36.
- Atkinson, A. C. (1985). *Plots, transformations, and regression: An introduction*

to graphical methods of diagnostic regression analysis. Oxford: Clarendon Press.

Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*, 27, 17–21.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way model. *Annals of Mathematical Statistics*, 25, 290–302.

Box, G. E. P., & Cox, D.R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, 26, 211–246.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis*. New York: Chapman and Hall.

Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician*, 35, 124–129.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, 74, 169–174.

Emerson, J. D. (1991). Introduction to transformation. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Fundamentals of exploratory analysis of variance* (pp. 365–400). New York: Wiley.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23, 524–532.

Judd, C. M., & McClelland, G. H. (1998). Measurement. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., pp. 180–232). New York: McGraw-Hill.

Judd, C. M., McClelland, G. H., & Culhane, S. E. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. *Annual Review of Psychology*, 46, 433–465.

Judd, C. M., McClelland, G. H., & Ryan, C.R. (2009). *Data analysis: A model comparison approach* (2nd ed.). New York: Psychology Press/Taylor & Francis Group.

- Madansky, A. (1988). *Prescriptions for working statisticians*. New York: Springer-Verlag.
- SAS Institute. (1989). *SAS/STAT user's guide* (Version 6, 4th ed.). Cary, NC: SAS Institute.
- Shapiro, S. S., & Wilk, M. B. (1968). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Tukey, J. W., Mosteller, F., & Hoaglin, D. C. (1991). Concepts and examples in analysis of variance. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Fundamentals of exploratory analysis of variance* (pp. 1–23). New York: Wiley.
- Velleman, P. F. (1997). The philosophical past and the digital future of data analysis: 375 years of philosophical guidance for software design on the occasion of John W. Tukey's 80th birthday. In D. R. Brillinger, L. T. Fernholz, & S. Morgenthaler (Eds.), *The practice of data analysis: Essays in honor of John W. Tukey* (pp. 317–337). Princeton, NJ: Princeton University Press.
- Velleman, P. F., & Welsch, R. E. (1981). Efficient computing of regression diagnostics. *American Statistician*, 35, 234–242.
- Wilcox, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology*, 38, 29–60.
- Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.
- Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, 55, 1–17.

* Direct correspondence to Gary McClelland, Department of Psychology, CB345, University of Colorado, Boulder, CO 80309–0345 (e-mail:

gary.mcclelland@colorado.edu).

Chapter twenty-four Missing Data Analysis

Gina L. Mazza and Craig K. Enders

Missing data are pervasive in almost all research involving quantitative methods, including social psychology research. For years, the standard practice was to exclude cases with missing scores and perform analyses using only the subsample of participants with complete data on a particular set of variables. In a 1999 report, the American Psychological Association's Task Force on Statistical Inference admonished this approach, stating that excluding cases with missing scores is “among the worst methods available for practical applications” (Wilkinson & Task Force on Statistical Inference, 1999, p. 598). Their rationale was simple: Missing data theory and an accumulation of methodological research demonstrate that excluding incomplete cases reduces statistical power and provides parameter estimates that are prone to substantial bias. Despite the Task Force's justifiably strong stance, literature reviews suggest that researchers still routinely employ subpar missing data handling methods (Bodner, 2006; Jeličić, Phelps, & Lerner, 2009; Peugh & Enders, 2004; Wood, White, & Thompson, 2004).

Several factors contribute to the disparity between what methodologists recommend and what researchers actually do. First, most general-use software packages exclude incomplete cases by default. Consequently, researchers may view this procedure as adequate when really it is not. Second, graduate-level statistics courses and textbooks rarely, if ever, address modern missing data handling methods. Rather, graduate students learn how to perform analyses with complete data sets (which they may rarely see in their own research); this tacitly reinforces the notion that default software settings lead to correct analytic practices. Third, much of the missing data literature is highly technical and not aimed at researchers looking to apply modern missing data handling methods to their research.

Methodologists have been interested in missing data issues for almost a century, particularly since the 1970s with the introduction of two so-called modern missing data handling methods: multiple imputation and maximum likelihood estimation (Beale & Little, 1975; Dempster, Laird, & Rubin, 1977;

Rubin, 1987). The literature currently recommends these approaches because they (1) invoke more realistic assumptions about the reasons for missing data, (2) improve the accuracy of parameter estimates, and (3) mitigate the loss of statistical power. Multiple imputation and maximum likelihood estimation are now available in many general-use statistical software packages, are easy to implement, and require relatively short computing times. Thus, researchers need to understand how to apply these methods to their own research. This chapter provides a nontechnical overview of missing data issues, with a particular emphasis on multiple imputation and maximum likelihood estimation. To facilitate the adoption of these procedures, we target this chapter at researchers who have graduate-level training in multiple regression analysis.

The organization of this chapter is as follows. First, we describe three missing data mechanisms that explain why data are missing. We then discuss planned missing data designs that can reduce respondent burden and lower the cost of data collection. Next, we describe traditional missing data handling methods commonly used by researchers, and we explain why these approaches produce biased parameter estimates and reduce the statistical power to detect effects. Having established the necessary background, we spend the majority of this chapter discussing multiple imputation and maximum likelihood estimation. Without getting too technical, we explain how these procedures work and use a small data set to demonstrate their application. We then discuss practical considerations that influence the choice between multiple imputation and maximum likelihood estimation. Finally, we briefly introduce a class of analysis models that are designed to address the difficult case of not missing at random data.

Missing Data Mechanisms

Missing data mechanisms describe how the probability of missing data on a variable Y relates to other variables in the analysis or to the values of Y itself. Rubin and colleagues (Little & Rubin, 2002; Rubin, 1976) classify missing data according to three mechanisms: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). These mechanisms serve as assumptions used by missing data handling methods; conceptually, they describe different reasons for missingness. When researchers perform a missing data analysis (even just deleting incomplete cases), they implicitly assume that data are missing according to one of these three mechanisms. Applying a procedure with incorrect assumptions usually results in biased parameter

estimates.

To illustrate the three missing data mechanisms, [Table 24.1](#) gives a data set with 20 participants. Suppose that a health psychologist collects the data in [Table 24.1](#) to investigate the effectiveness of a weight loss program. Participants first complete a questionnaire regarding their self-efficacy to lose weight along with other questionnaires related to diet, exercise, weight, and personality. They then participate in a 12-week weight loss program and are weighed at the end of 12 weeks to determine their body mass index (BMI). The “Complete” column represents the self-efficacy scores and BMIs we would have observed had the data been complete. We then deleted BMIs from the hypothetical complete data set according to an MCAR, MAR, or NMAR mechanism. We compare the MCAR, MAR, and NMAR data sets to the hypothetical complete data set throughout this section to demonstrate how the missing data mechanisms affect our analyses.

Table 24.1. BMI with MCAR, MAR, and NMAR Data

Self-Efficacy	Complete	MCAR		MAR		NMAR	
		Indicator	BMI	Indicator	BMI	Indicator	BMI
10.0	38.1	1	38.1	0	—	0	—
16.0	35.4	1	35.4	0	—	1	35.4
16.0	34.0	0	—	0	—	1	34.0
17.0	36.8	1	36.8	0	—	0	—
19.0	31.3	1	31.3	0	—	1	31.3
23.0	38.2	1	38.2	1	38.2	0	—
24.0	35.9	0	—	1	35.9	0	—
26.0	34.7	1	34.7	1	34.7	1	34.7
26.0	32.6	1	32.6	1	32.6	1	32.6
28.0	29.8	0	—	1	29.8	1	29.8
31.0	38.2	1	38.2	1	38.2	0	—
37.0	33.5	1	33.5	1	33.5	1	33.5
37.0	31.2	0	—	1	31.2	1	31.2
38.0	27.0	1	27.0	1	27.0	1	27.0
40.0	34.9	1	34.9	1	34.9	1	34.9
44.0	34.4	1	34.4	1	34.4	1	34.4
45.0	28.4	1	28.4	1	28.4	1	28.4
47.0	29.1	1	29.1	1	29.1	1	29.1
50.0	25.6	0	—	1	25.6	1	25.6
66.0	32.3	1	32.3	1	32.3	1	32.3

Missing Completely at Random (MCAR)

The MCAR mechanism states that the probability of missing data on Y is unrelated to other variables in the analysis and to the values of Y itself. The

complete cases are a random subsample of the hypothetical complete data set, and are thus still representative of the population. Returning to [Table 24.1](#), an MCAR mechanism requires that neither self-efficacy nor the values of BMI itself are related to the probability of having a missing BMI. More generally, the MCAR mechanism requires that *none* of the other variables in the data set (e.g., diet, exercise, weight, personality) predict whether or not a participant has a missing BMI. For example, an MCAR mechanism may include a participant not being weighed because the scale was malfunctioning or a participant's BMI record being misplaced prior to data entry. We emulated an MCAR mechanism in [Table 24.1](#) by randomly deleting five participants' BMIs.

To some extent, the data provide evidence for or against an MCAR mechanism. Because the MCAR mechanism implies that the probability of missing data on Y is unrelated to other variables in the analysis, participants with observed scores on Y and participants with missing scores on Y should not differ on other variables. Returning to the MCAR data set in [Table 24.1](#), we created a binary missing data indicator that uses codes of 1 and 0 to denote cases with observed and missing BMIs, respectively. If the MCAR mechanism is plausible, the missing data indicator should not be related to other variables in the data set. When we split participants in the MCAR data set by the missing data indicator, the average self-efficacy score for participants with observed BMIs is 32.00 and the average self-efficacy score for participants with missing BMIs is 32.33. An independent t -test revealed that these means are equivalent, $t(18) = 0.07$, $p = .95$, which supports the MCAR mechanism (Dixon, [1988](#)). To fully evaluate the MCAR mechanism, we would perform t -tests for all other variables in the data set.

In the previous example, the nonsignificant mean difference provides some evidence for the MCAR mechanism. However, note that (1) we can never prove the null hypothesis using a nonsignificant t -test; (2) the subgroups often have very unequal sample sizes (e.g., only five participants have missing BMIs), which severely reduces the statistical power to detect mean differences; and (3) the MAR and NMAR mechanisms (which we discuss later in this section) can produce subgroups with equal means (e.g., if scores are systematically missing from the upper and lower tails of the BMI distribution). We can compute effect sizes such as standardized mean differences (e.g., Cohen's d) to mitigate the loss of statistical power to detect mean differences, but effect sizes are also affected by unequal sample sizes. Thus, researchers cannot automatically assume an MCAR mechanism based on a nonsignificant t -test or an effect size close to zero. As we discuss in the next section, the default procedures in many statistical

software packages (e.g., deleting incomplete cases) assume an MCAR mechanism. However, the MCAR mechanism is a very strict assumption that is implausible for most data sets.

Missing at Random (MAR)

The MAR mechanism states that the probability of missing data on Y is completely explained by other variables in the analysis and is unrelated to the values of Y itself. Said differently, after controlling for other variables in the analysis, missingness is unrelated to the values of Y itself. Saying that these data are MAR is misleading because the data are *not* randomly missing, at least not in the flip-of-a-coin sense of randomness. Rather, the MAR mechanism describes systematic missingness, such that the probability of missing data on Y is completely explained by one or more other variables in the analysis. Returning to our earlier example, suppose that participants with low self-efficacy to lose weight drop out before completing the weight loss program and thus are not weighed at the end of 12 weeks. This example qualifies as an MAR mechanism if missingness is unrelated to the values of BMI itself after controlling for participants' self-efficacy scores (i.e., for two participants with the same self-efficacy score, the likelihood of having a missing BMI is completely random). We emulated an MAR mechanism in [Table 24.1](#) by deleting BMIs for the five participants with the lowest self-efficacy scores.

Unlike the MCAR mechanism, the data do not provide evidence for or against an MAR mechanism. We cannot demonstrate that the probability of missing data on Y is unrelated to the values of Y itself after controlling for other variables without knowing what the scores on Y would have been had the data been complete. Returning to the MAR data set in [Table 24.1](#), we created a binary missing data indicator that uses codes of 1 and 0 to denote cases with observed and missing BMIs, respectively. When we split participants in the MAR data set by the missing data indicator, the average self-efficacy score for participants with observed BMIs is 37.47 and the average self-efficacy score for participants with missing BMIs is 15.60. An independent t -test indicated that this mean difference is significant, $t(18) = 4.02$, $p = .001$. This large mean difference makes the MCAR mechanism implausible because missingness is systematically related to self-efficacy (i.e., participants with low self-efficacy scores are more likely to have missing BMIs). However, the MAR mechanism requires that the probability of having a missing BMI is unrelated to the values of BMI itself after controlling for participants' self-efficacy scores. Unfortunately, we cannot

evaluate this proposition when the BMIs are missing. Thus, any analysis that requires an MAR mechanism ultimately relies on an untestable assumption, although this assumption is much less strict than an MCAR mechanism. As we discuss later in this chapter, multiple imputation and maximum likelihood estimation assume an MAR mechanism, but they also provide unbiased parameter estimates with an MCAR mechanism .

Not Missing at Random (NMAR)

The NMAR mechanism states that the probability of missing data on Y is related to the values of Y itself, even after controlling for other variables in the analysis. Returning to our earlier example, suppose that participants with very high BMIs refuse to be weighed at the end of 12 weeks. Alternatively, suppose that participants with very high BMIs drop out because the weight loss program is too strenuous or because they believe that it is not working for them. This example qualifies as an NMAR mechanism if BMI values predict missingness even after controlling for other variables such as self-efficacy (e.g., for two participants with the same self-efficacy score, the participant with the higher BMI is more likely to quit the weight loss program). We emulated an NMAR mechanism in [Table 24.1](#) by deleting the five highest BMIs.

Like the MAR mechanism, the data do not provide evidence for or against an NMAR mechanism. We cannot demonstrate that the probability of missing data on Y is related to the values of Y itself after controlling for other variables without knowing what the scores on Y would have been had the data been complete. Returning to the NMAR data set in [Table 24.1](#), we created a binary missing data indicator that uses codes of 1 and 0 to denote cases with observed and missing BMIs, respectively. When we split participants in the NMAR data set by the missing data indicator, the average self-efficacy score for participants with observed BMIs is 35.67 and the average self-efficacy score for participants with missing BMIs is 21.00. An independent t -test indicated that this mean difference is significant, $t(18) = 2.21$, $p = .04$. This large mean difference rules out an MCAR mechanism, but we cannot verify that the mechanism is NMAR because we cannot establish that the values of BMI itself predict missingness above and beyond other variables such as self-efficacy (doing so requires the would-be BMIs).

Summary

Here, we present overly simplistic examples to illustrate Rubin's missing data

mechanisms. In practice, the causes of missingness are often complex. Multiple variables may predict missingness, and the reasons for missingness may vary across participants or across incomplete variables. As a further complication, researchers may not have measured the causes of missingness. Despite these complications, researchers must adopt one of the three mechanisms when performing a missing data analysis. The missing data literature provides virtually no support for MCAR-based missing data handling methods (e.g., excluding incomplete cases). Consequently, we usually have to decide between missing data handling methods that assume an MAR or an NMAR mechanism. Our previous examples demonstrate that inspecting the data does not inform this decision because the MAR and NMAR mechanisms both make propositions about *unobserved* scores (e.g., the would-be BMIs in our examples). The majority of this chapter is devoted to two MAR-based analyses: multiple imputation and maximum likelihood. These modern missing data handling methods are easy to use and are readily available in general-use statistical software packages. Further, the MAR assumption is often quite plausible for social psychology research. At the end of this chapter, we briefly introduce two missing data handling methods that assume an NMAR mechanism (the selection model and the pattern mixture model). We recommend using these procedures with caution because they require strict (and often unrealistic) assumptions and may produce biased estimates (Enders, 2010; Schafer & Graham, 2002). Nevertheless, it is important to raise awareness of NMAR-based analyses.

Planned Missing Data Designs

In the previous section, we provided examples where the reasons for missing data were outside of the researcher's control. Methodologists have developed planned missing data designs that purposefully introduce missing scores in order to reduce respondent burden (e.g., by reducing the number of questionnaires that participants need to complete) and expenses related to data collection. Researchers are often reluctant to introduce planned missing data into their research designs because they incorrectly assume that all missing scores are harmful. However, because the missing data are under the researcher's control, the resulting mechanism is MCAR or MAR, by definition. Consequently, we can use multiple imputation or maximum likelihood estimation to obtain unbiased parameter estimates with planned missing data designs. Interestingly, the loss of statistical power is often much smaller than you might expect. The possibility of reducing respondent burden without incurring bias or a substantial reduction in

statistical power makes planned missing data designs a valuable but underutilized tool for social psychology research.

Actually, you are probably already familiar with planned missing data designs. In a randomized research design that assigns participants to either an experimental or control condition, participants hypothetically have scores from both conditions even though they only provide a score from their assigned condition. A participant's hypothetical score from the unassigned condition, referred to as the participant's counterfactual score, is MCAR because the researcher randomly assigns participants to conditions. As another example, the electronic GRE uses a planned missing data design with an MAR mechanism. Test takers get different subsets of items, and some test takers get fewer items than others. An algorithm decides which item to give the test taker, depending on whether or not he or she answered the previous item correctly, so the mechanism is MAR because missingness on certain items is solely determined by the test taker's responses to previous items. This process is explained more fully by Widaman and Grimm (Chapter 20 in this volume).

In this section, we describe three planned missing data designs that create MCAR data: the two-method measurement design, the three-form design, and a variation of the three-form design for repeated measures and longitudinal research designs. Because planned missing data designs are highly useful for reducing respondent burden and expenses related to data collection, they will likely gain popularity as researchers become more familiar with them. Refer to Graham, Hofer, and MacKinnon (1996), Graham, Taylor, and Cumsille (2001), and Graham, Taylor, Olchowski, and Cumsille (2006) for a more comprehensive explanation of planned missing data designs.

Two-Method Measurement Design

The two-method measurement design uses (1) a cheap but less valid measure and (2) an expensive but more valid (e.g., “gold standard”) measure of a single construct. The two-method measurement design, in conjunction with modern missing data handling methods, allows researchers to administer the more expensive measure to a random subsample of the participants but analyze data from all participants without incurring bias. Returning to our earlier example, suppose that the health psychologist plans to measure participants' body fat percentages. A BOD POD®, which measures body fat percentage using air displacement, provides highly accurate estimates but is too expensive for her grant budget. However, she can cheaply measure participants' body fat

percentages by taking skinfold measurements with a set of calipers. In this example, a two-method measurement design involves taking skinfold measurements from all the participants and using the more expensive BOD POD on a random subsample of the participants. BOD POD scores are MCAR because she randomly assigns participants to either have BOD POD scores or not. As we discuss later in this chapter, multiple imputation and maximum likelihood estimation allow us to analyze data from all the participants rather than from just the subsample of participants with BOD POD scores. For example, we estimate the correlation between BOD POD scores and self-efficacy scores using data from all participants. Conceptually, multiple imputation and maximum likelihood estimation use information from the skinfold measurements to estimate the association between BOD POD scores and self-efficacy scores.

Three-Form Design

When researchers measure constructs using questionnaires, the number of items can quickly add up and create substantial respondent burden. Researchers may be limited by time (e.g., students in introductory psychology courses are only required to participate for one hour to receive credit), cost (e.g., participants are paid hourly), or participant engagement (e.g., participants get bored and rush through the items). To reduce the number of items, researchers usually cut questionnaires or items within questionnaires. However, they should also consider using the three-form design that uses the entire questionnaire battery but distributes subsets of items or questionnaires randomly across participants (Graham et al., 1996; Graham et al., 2006).

The three-form design divides the items into four subsets (X, A, B, C) and allocates these subsets of items to three questionnaire forms. All three questionnaire forms contain the X subset and two other subsets (i.e., A and B, A and C, B and C), such that each questionnaire form is missing one subset (i.e., A, B, or C). Table 24.2 shows the three-form design. Returning to our earlier example, suppose that the health psychologist wants to administer four 20-item questionnaires, but her participants only have time to answer 60 questions. Using the three-form design, she assigns 20 items each to the X, A, B, and C subsets and creates three questionnaire forms. She then randomly assigns participants to receive one of the three questionnaire forms. Each participant only answers 60 questions (i.e., the X subset and two other subsets), but the health psychologist can use all 80 items in her analyses. Note that the subsets do not need to have an

equal number of items. We can include multiple questionnaires in each subset and/or divide items from a single questionnaire across subsets .

Table 24.2. Three-Form Design

	Item Subsets			
Form	X	A	B	C
1	O	M	O	O
2	O	O	M	O
3	O	O	O	M

Note: Os denote observed data and Ms denote missing data.

Planned Missing Data Designs for Repeated Measures

Respondent burden and expenses related to data collection are also practical considerations for repeated measures and longitudinal research designs with multiple waves of data collection. Returning to our earlier example, suppose that the health psychologist wants to track her participants' progress once a week during the 12-week weight loss program, but asking her participants to come to her office 12 times creates substantial respondent burden. Graham *et al.* (2001) describe variations of the three-form design for repeated measures and longitudinal research designs with multiple waves of data collection. Participants are split into random subsamples, and each subsample misses a wave (or waves) of data collection. Interestingly, Graham *et al.* (2001) demonstrate that planned missing data designs can offer more statistical power to detect effects than complete data designs that use the same number of data points (i.e., observations within a data set). Although this finding needs clarification, it appears that collecting incomplete data from a larger sample is more advantageous for statistical power than collecting complete data from a smaller sample.

Traditional Missing Data Handling Methods

As stated earlier, researchers often rely on traditional missing data handling methods even though methodologists recommend using either multiple imputation or maximum likelihood estimation to address missing data. Two common strategies – deleting incomplete cases (i.e., deletion methods) or filling in a single score for each missing score (i.e., single imputation) – often assume an MCAR mechanism and produce biased parameter estimates when this assumption is violated. Even when the MCAR assumption is met, traditional missing data handling methods reduce the statistical power to detect effects and some (e.g., filling in the missing scores with the mean of the observed scores) provide biased parameter estimates. To illustrate, we apply the traditional missing data handling methods discussed in this section to the MAR data set in [Table 24.1](#). Recall that participants with low self-efficacy to lose weight drop out before completing the weight loss program and thus have missing BMIs. [Figure 24.1](#) shows a scatterplot of the hypothetical complete data set in [Table 24.1](#), which we later compare to the scatterplots from traditional missing data handling methods to see how well they reproduce features of the complete data. Although we discuss the deletion methods and single imputation methods you are most likely to see in published research articles and statistical software packages, a variety of sources provide information on other traditional missing data handling methods (Allison, 2002; Enders, 2010; Little & Rubin, 2002; Schafer & Graham, 2002; van Buuren, 2012).

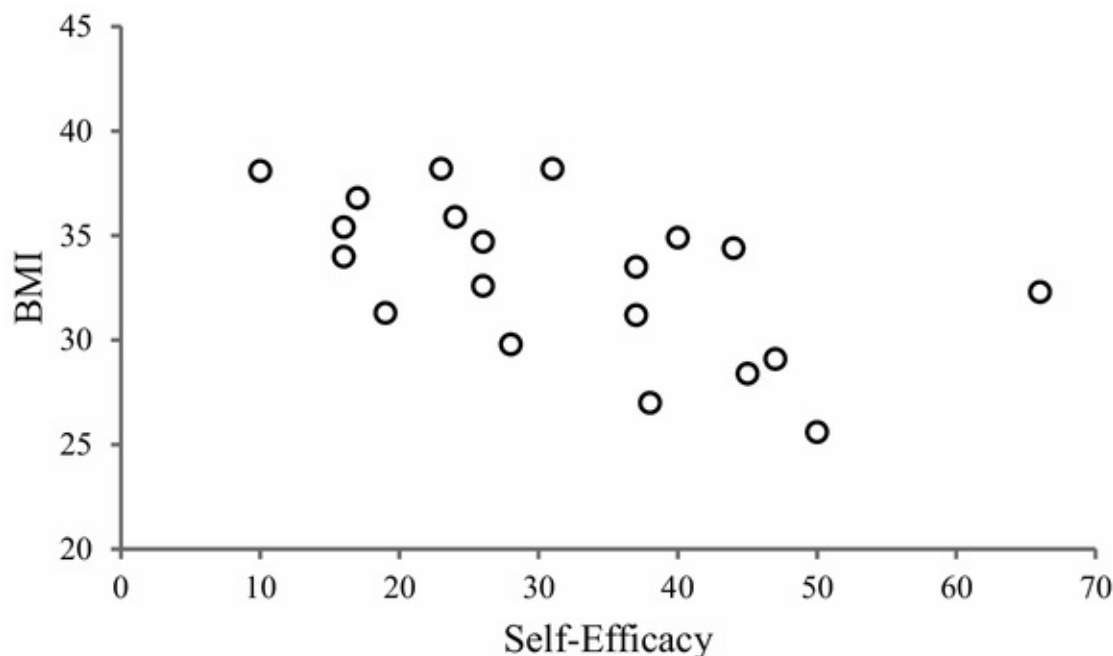


Figure 24.1. Scatterplot of hypothetical complete data set in [Table 24.1](#).

Deletion Methods

Listwise deletion (also called complete-case analysis or casewise deletion) removes cases (i.e., participants) with one or more missing scores, so only cases with complete data remain in the data set. Pairwise deletion (also called available-case analysis) removes cases on an analysis-by-analysis basis (e.g., computing a series of correlations using the cases with complete data on each variable pair). Unlike with listwise deletion, analyses performed with pairwise deletion use varying sample sizes depending on which variables are involved. Statistical software packages make listwise and pairwise deletion highly convenient, but removing cases with missing scores reduces the sample size and thus the statistical power to detect effects. Deletion methods also provide biased parameter estimates when the MCAR assumption is violated.

Consider the MAR data set in [Table 24.1](#). [Figure 24.2](#) shows a scatterplot of the MAR data set in [Table 24.1](#) after deleting cases with missing BMIs. Self-efficacy and BMI are negatively correlated, so removing cases from the lower tail of the self-efficacy distribution also removes cases from the upper tail of the BMI distribution. Listwise deletion produces a self-efficacy mean that is too high ($M = 37.47$ versus 32.00 in the hypothetical complete data set) and a BMI mean that is too low ($M = 32.39$ versus 33.07), as shown in [Table 24.3](#). Because systematic missingness (e.g., an MAR mechanism) often restricts variability in a data set, listwise deletion tends to attenuate estimates of variation (e.g., standard deviations) and association (e.g., correlations). Such is the case in [Table 24.3](#), where the correlation drops from $-.56$ to $-.48$. Wilkinson and the APA Task Force on Statistical Inference (1999) describe listwise and pairwise deletion as among the worst missing data handling methods because these approaches require the unrealistic MCAR assumption and yield biased parameter estimates when the mechanism is MAR or NMAR. Nevertheless, deletion methods are often reported in published research articles more than a decade later.

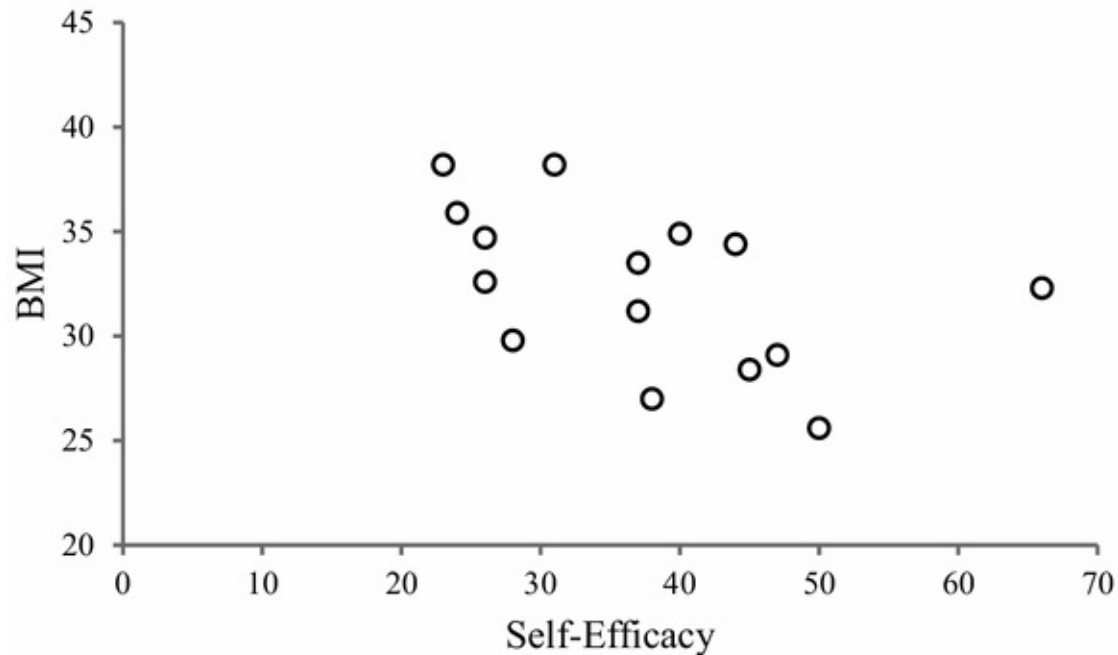


Figure 24.2. Listwise deletion scatterplot.

Table 24.3. *BMI Parameter Estimates from Traditional Missing Data Handling Methods*

	Mean	Standard Deviation	Correlation
Complete	33.07	3.72	-.56
Listwise Deletion	32.39	3.84	-.48
Mean Imputation	32.39	3.30	-.35
Regression Imputation	33.24	3.64	-.61
Stochastic Regression Imputation	33.00	3.87	-.53

Mean Imputation and Averaging the Available Items

Single imputation methods impute (i.e., fill in) a single score for each missing

score. In the remainder of this section, we introduce four single imputation methods: mean imputation, averaging the available items, regression imputation, and stochastic regression imputation. Mean imputation (also called mean substitution and unconditional mean imputation) imputes the missing scores with the mean of the observed scores. Because adding scores to the center of a variable's distribution reduces the variability of the data set, mean imputation attenuates estimates of variation and association. [Figure 24.3](#) shows a scatterplot of the MAR data set in [Table 24.1](#) after using mean imputation. We averaged the 15 observed BMIs and replaced the five missing BMIs with $M = 32.39$. As you can see in [Figure 24.3](#), the imputed BMIs fall on a straight line with a slope of zero, meaning the correlation between self-efficacy and BMI is $r = 0$ for cases with imputed BMIs. Because mean imputation merges the uncorrelated imputed BMIs with the observed BMIs, the correlation between self-efficacy and BMI is much closer to zero in the mean imputation data set than in the hypothetical complete data set, as shown in [Table 24.3](#). Imputing the missing scores with the mean also reduces the variability in the data set (see the standard deviations in [Table 24.3](#)). Like listwise deletion, mean imputation yields a BMI mean that is too low because data are systematically missing from the upper tail of the BMI distribution. Some researchers regard mean imputation as a conservative missing data handling method because it underestimates associations in the population. However, the bias is usually so severe that conservatism is simply an unacceptable rationale for implementing mean imputation. Simulation studies suggest that mean imputation yields greater bias than other missing data handling methods (Brown, [1994](#); Enders & Bandalos, [2001](#); Olinsky, Chen, & Harlow, [2003](#)), so researchers should never rely on mean imputation to address missing data.

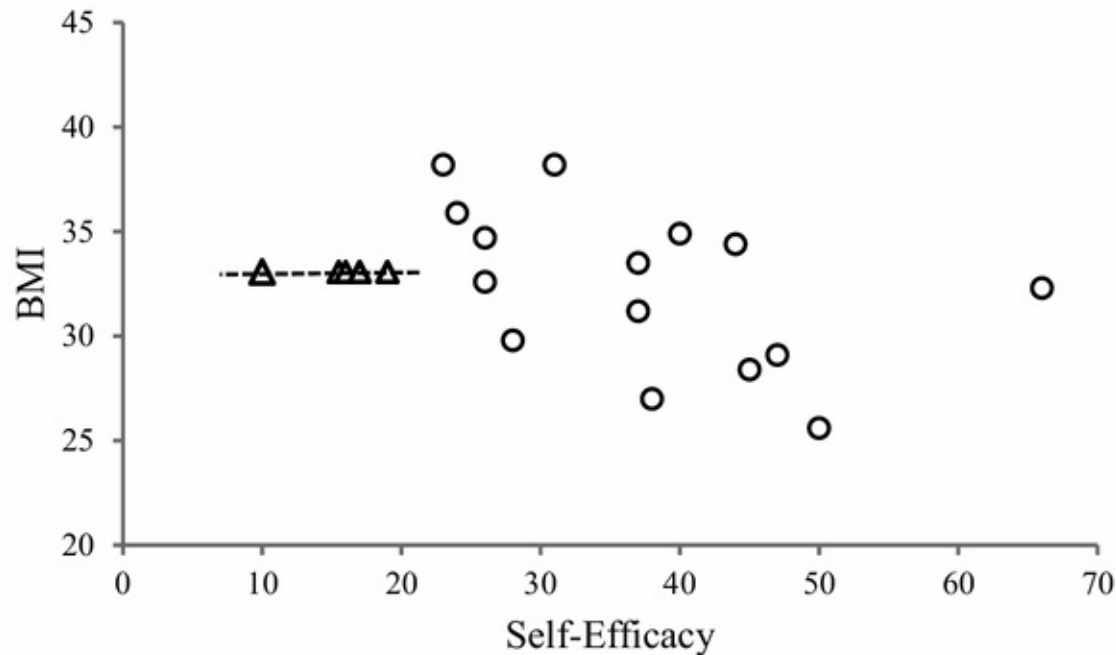


Figure 24.3. Mean imputation scatterplot. Circular points represent observed scores and triangular points represent imputed scores.

Averaging the available items is another version of mean imputation used when computing scale scores. Researchers often use questionnaires to measure constructs and compute scale scores by summing or averaging the items that measure a single construct. When participants have observed scores on some, but not all, of the items that measure a single construct, researchers often compute the scale scores by averaging the available items. Averaging the available items is equivalent to imputing each participant's missing scores with the mean of his or her observed scores, which is why it is also referred to as person mean imputation. To illustrate, [Table 24.4](#) shows two rows of data for four participants. The first row shows the observed scores along with the average of the available items in the scale score column. The second row replaces the missing scores with the average of the available items from the first row. Notice that the scale scores are identical in both rows, which demonstrates that averaging the available items is another version of mean imputation. Relatively few simulation studies have investigated the performance of averaging the available items, but our own simulations suggest that it produces biased parameter estimates with an MAR mechanism. Implementing this approach is perhaps most defensible when the items have uniform means and correlations because a participant's average score may be fairly representative of the item responses. However, psychometricians often recommend writing items that elicit

variability among participants (Clark & Watson, 1995), in which case averaging the available items would likely produce biased parameter estimates. Although averaging the available items is commonly used to address item-level nonresponse, multiple imputation (which we discuss later in this chapter) is a much better solution.

Table 24.4. Computing Scale Scores by Averaging the Available Items or Using Person Mean Imputation

Participant	Method	Item 1	Item 2	Item 3	Item 4	Scale Score
1	AAI	5.0	6.0	–	4.0	5.0
	PMI	5.0	6.0	5.0	4.0	5.0
2	AAI	2.0	–	1.0	–	1.5
	PMI	2.0	1.5	1.0	1.5	1.5
3	AAI	–	7.0	7.0	6.0	6.7
	PMI	6.7	7.0	7.0	6.0	6.7
4	AAI	4.0	5.0	–	–	4.5
	PMI	4.0	5.0	4.5	4.5	4.5

Note: AAI = averaging the available items. PMI = person mean imputation.

Regression Imputation

Regression imputation (also called conditional mean imputation) imputes the missing scores with predicted scores from a regression equation. First we estimate a regression equation that predicts a variable with missing data from variables with complete data. Then we use the predicted scores from this regression equation to impute the missing scores. In the MAR data set in [Table 24.1](#), we use the 15 cases with complete data to estimate the following regression equation that predicts participants' BMIs from their self-efficacy scores:

$$\text{BMI}_{\text{imputed}} = 38.25 - 0.16 (\text{Self-Efficacy}) \quad (24.1)$$

We compute the predicted BMIs by substituting participants' self-efficacy scores into the regression equation and impute the missing BMIs with the predicted BMIs.

[Figure 24.4](#) shows a scatterplot of the MAR data set in [Table 24.1](#) after using

regression imputation. Notice that the imputed BMIs fall directly on the regression line (i.e., self-efficacy and BMI are perfectly correlated for cases with missing BMIs), which is unlikely had the data been complete. As you can see in [Table 24.3](#), regression imputation reduces the variability in the data set, which attenuates estimates of variation. Because the correlation between self-efficacy and BMI is $r = -1$ for cases with imputed BMIs, regression imputation also overestimates the correlation between self-efficacy and BMI. Regression imputation provides unbiased estimates of the means with an MCAR or MAR mechanism, but this is not a compelling reason to use it because we are usually interested in estimating associations among variables.

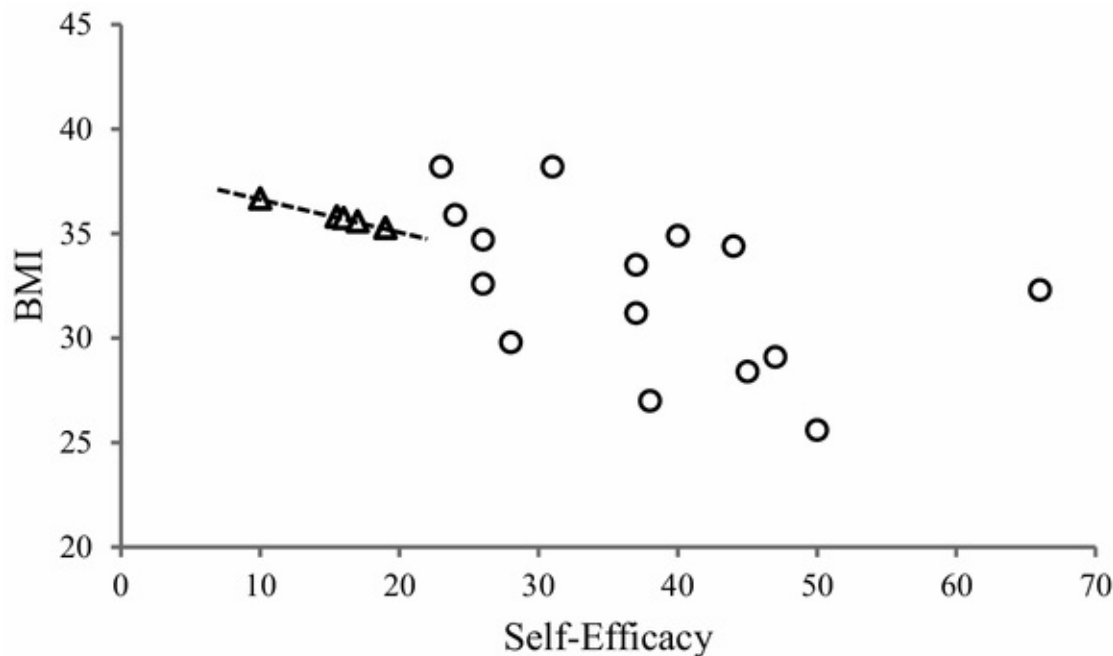


Figure 24.4. Regression imputation scatterplot. Circular points represent observed scores and triangular points represent imputed scores.

Stochastic Regression Imputation

Regression imputation provides biased parameter estimates because the imputed scores lack variability, so stochastic regression imputation adds normally distributed residuals (i.e., random noise) to the imputed scores. Like regression imputation, we estimate a regression equation that predicts a variable with missing data from variables with complete data and then compute the predicted scores (see Equation 24.1). Unlike regression imputation, we then add a residual randomly drawn from a normal distribution with a mean of zero and a variance equal to the residual variance from the regression of BMI on self-efficacy. The

imputed scores are the sums of the predicted scores and the normally distributed residuals. Adding normally distributed residuals to the predicted scores restores variability to the data set and eliminates the biases associated with regression imputation.

Figure 24.5 shows a scatterplot of the MAR data set in Table 24.1 after using stochastic regression imputation. Notice that the imputed BMIs do not fall directly on the regression line. Because adding normally distributed residuals (two of which are denoted as arrows in Figure 24.5) to the predicted BMIs restores variability to the data set, the scatterplot from the stochastic regression imputation data set in Figure 24.5 resembles the scatterplot from the hypothetical complete data set in Figure 24.1. As you can see in Table 24.3, the BMI mean and the correlation between self-efficacy and BMI are very close to the parameter estimates from the hypothetical complete data set. Unlike other traditional missing data handling methods, stochastic regression imputation provides unbiased parameter estimates with an MCAR or MAR mechanism. However, analyzing an imputed data set as though it were complete yields standard errors that are too low because statistical software packages cannot account for the fact that the imputed scores are just one set of plausible replacement scores (i.e., our best guesses about the missing scores). Because the standard errors do not account for imputation uncertainty, significance tests based on stochastic regression imputation suffer from excessive Type I error rates. Although we can use stochastic regression imputation for analyses that do not require significance tests (e.g., factor analyses and reliability analyses), multiple imputation and maximum likelihood estimation offer greater flexibility because they yield accurate standard errors .

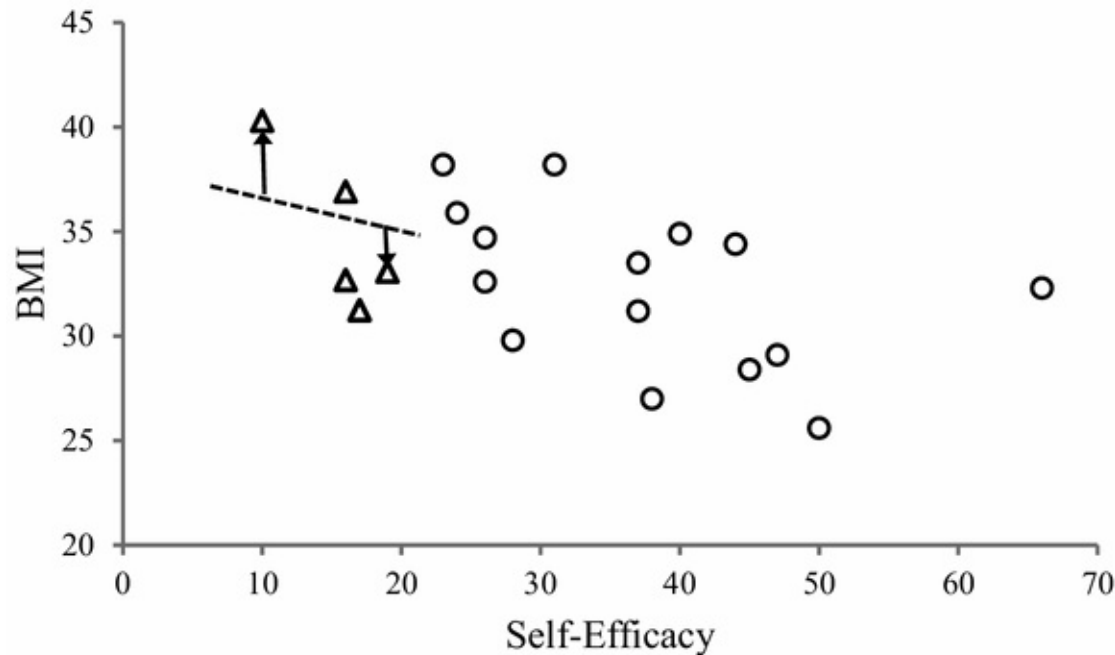


Figure 24.5. Stochastic regression imputation scatterplot. Circular points represent observed scores and triangular points represent imputed scores. The arrows depict normally distributed residuals from the regression line.

Multiple Imputation

Multiple imputation is one of the two modern missing data handling methods that methodologists currently recommend because it provides accurate parameter estimates with an MAR mechanism. Multiple imputation consists of three phases: the imputation phase, the analysis phase, and the pooling phase. Unlike single imputation methods, multiple imputation creates multiple copies of the data set and imputes each copy with different scores. We then analyze the imputed data sets as though they were complete data sets. Finally, we pool the parameter estimates across the imputed data sets, which yields a single set of parameter estimates.

Imputation Phase

The imputation phase of multiple imputation creates multiple (preferably 20 or more) copies of the data set and imputes each copy with different scores (Graham, Olchowski, & Gilreath, 2007). Methodologists have proposed several algorithms for the imputation phase, but we focus on Schafer's (1997) data augmentation algorithm, which assumes multivariate normality for the incomplete variables. The data augmentation algorithm consists of two iterative

(i.e., repeated) steps: the imputation step (I-step) and the posterior step (P-step). Procedurally, the I-step is equivalent to stochastic regression imputation. To impute the missing scores, we compute predicted scores from a regression equation that predicts a variable with missing data from variables with complete data and then add normally distributed residuals.

The imputed BMIs from the I-step represent one set of plausible BMIs for participants with missing BMIs. To generate another set of plausible BMIs, we need to use a different regression equation to predict the missing BMIs. Following the I-step, the P-step computes the mean vector (i.e., the mean of each variable) and the covariance matrix (i.e., the variance of each variable and the covariance between each pair of variables) for the imputed data set. Using the resulting parameter estimates, the P-step defines a posterior distribution (the Bayesian analog of a sampling distribution) of plausible mean vectors and covariance matrices and uses a computer simulation technique to sample an alternate mean vector and covariance matrix from each distribution. The algorithm then uses the alternate mean vector and covariance matrix to solve for the regression coefficients in the subsequent I-step. Conceptually, randomly drawing an alternate mean vector and covariance matrix from the posterior distribution is akin to adding random noise to the regression coefficients in the regression equation from the preceding I-step. Because we compute predicted scores from a different regression equation, the imputed data set from the second I-step differs from the imputed data set from the first I-step, the imputed data set from the third I-step differs from the imputed data set from the second I-step, and so on. Iterating through the I-step and P-step yields multiple copies of the data set with different imputed scores. Thus, multiple imputation is intuitively appealing because it does not rely on a single regression equation to impute the missing scores. Rather, it generates imputed scores from a random sample of plausible population regression equations.

Using the MAR data set in [Table 24.1](#), we generated four imputed data sets with the data augmentation algorithm. [Table 24.5](#) shows the four imputed data sets. Methodologists recommend using 20 or more imputed data sets (Graham et al., 2007), but we used four in this example so that we could work through the calculations involved in multiple imputation. As you can see in [Table 24.5](#), the imputed scores differ across imputed data sets but the observed scores remain constant. Again, the imputed scores differ across imputed data sets because (1) we compute predicted scores from a different regression equation for each imputed data set, and (2) we add random noise to the predicted scores to generate the imputed scores.

Table 24.5. Imputed Data Sets from Multiple Imputation

Self-Efficacy	Complete	Observed	Imputed BMIs			
			Data Set 1	Data Set 2	Data Set 3	Data Set 4
10.0	38.1	–	40.3	33.0	37.6	31.7
16.0	35.4	–	34.5	35.1	34.8	40.9
16.0	34.0	–	40.1	35.1	41.6	35.6
17.0	36.8	–	35.7	31.8	35.0	27.5
19.0	31.3	–	34.3	34.3	38.2	35.8
23.0	38.2	38.2	38.2	38.2	38.2	38.2
24.0	35.9	35.9	35.9	35.9	35.9	35.9
26.0	34.7	34.7	34.7	34.7	34.7	34.7
26.0	32.6	32.6	32.6	32.6	32.6	32.6
28.0	29.8	29.8	29.8	29.8	29.8	29.8
31.0	38.2	38.2	38.2	38.2	38.2	38.2
37.0	33.5	33.5	33.5	33.5	33.5	33.5
37.0	31.2	31.2	31.2	31.2	31.2	31.2
38.0	27.0	27.0	27.0	27.0	27.0	27.0
40.0	34.9	34.9	34.9	34.9	34.9	34.9
44.0	34.4	34.4	34.4	34.4	34.4	34.4
45.0	28.4	28.4	28.4	28.4	28.4	28.4
47.0	29.1	29.1	29.1	29.1	29.1	29.1
50.0	25.6	25.6	25.6	25.6	25.6	25.6
66.0	32.3	32.3	32.3	32.3	32.3	32.3

Serial dependence of imputed data sets. Although not obvious, imputed scores from successive iterations of the I-step and P-step are serially dependent (i.e., correlated). This dependency arises because the imputed scores from the I-step influence the parameter estimates in the subsequent P-step, and the parameter estimates from the P-step influence the imputed scores in the subsequent I-step. To eliminate unwanted correlations among the imputed scores, we iterate through the I-step and P-step for hundreds or even thousands of cycles and save the imputed data sets at regular intervals. We specify a burn-in period (i.e., the number of iterations prior to saving the first imputed data set) and a between-imputation interval (i.e., the number of iterations that separate two saved imputed data sets). For example, to generate the four imputed data sets in [Table 24.5](#), we allowed the data augmentation algorithm to cycle for 500 iterations before saving the first imputed data set (the burn-in period), and then we saved an imputed data set every 200 iterations thereafter (the between-imputation interval). Separating the imputed data sets by 200 iterations ensures that the imputed scores are a random sample drawn from a distribution of plausible replacement scores. To set the between-imputation interval, we used numeric and graphical convergence diagnostic procedures provided by a

general-use statistical software package. A variety of sources describe these diagnostics in more detail (Enders, 2010; Schafer, 1997; Schafer & Olsen, 1998; van Buuren, 2012) .

Selecting variables for the imputation phase. At a minimum, the imputation regression equation must include all the variables and interactions used in subsequent analyses. Omitting a variable or interaction during the imputation phase attenuates its associations with other variables during the analysis phase. Methodologists further recommend an inclusive analysis strategy that adds auxiliary variables to the imputation regression equation (Collins, Schafer, & Kam, 2001). Auxiliary variables are often ancillary to the substantive research questions but are important for missing data handling because they predict missingness or correlate with the incomplete variable.

Auxiliary variables that predict missingness help us meet the MAR assumption and reduce bias. We can identify auxiliary variables using independent *t*-tests to examine mean differences between participants with observed and missing scores (recall that we demonstrated this procedure in the Missing Data Mechanisms section). Large mean differences suggest systematic differences between participants with observed and missing scores, so including these variables as auxiliary variables can reduce nonresponse bias. Returning to the MAR data set in Table 24.1, participants with low self-efficacy to lose weight drop out before completing the weight loss program and thus have missing BMIs. In this example, self-efficacy is of substantive interest for our analyses and already would have been included in the imputation regression equation. However, even if self-efficacy were not of substantive interest for our analyses, we need to include it in the imputation regression equation as an auxiliary variable because it predicts missingness (we know this because cases with observed and missing BMIs have different self-efficacy means). In fact, auxiliary variables are often ancillary to the substantive research questions. As another example of auxiliary variables, suppose that participants who live farther away from the health psychologist's office drop out before completing the weight loss program and thus have missing BMIs. Even though distance from the health psychologist's office is not of substantive interest, we need to include it as an auxiliary variable to meet the MAR assumption. Although the MAR mechanism requires that variables in the analysis completely explain the probability of missing data, empirical studies suggest that omitting a cause of missingness may not introduce substantial bias, particularly if the variable's partial correlation with the incomplete variable is less than .40 and the missing data rate is less than 25% (Collins et al., 2001). In practice, this means that we

can often approximate the MAR assumption by incorporating a set of auxiliary variables into the imputation phase.

Auxiliary variables that correlate with an incomplete variable help us more accurately predict the missing scores and mitigate the loss of statistical power that results from missing data. Returning to our earlier example, suppose that the health psychologist measured participants' body fat percentages using skinfold measurements. Body fat percentage correlates with BMI, so including body fat percentage as an auxiliary variable generates more accurate imputed scores and thus increases the statistical power to detect effects. Methodologists recommend including as many auxiliary variables as possible, especially with multiple imputation (Rubin, 1996). The largest increases in statistical power occur when auxiliary variables correlate with the incomplete variable at $r = .40$ or higher (Collins et al., 2001). It is important to note that including auxiliary variables in the imputation phase does not increase the likelihood of spurious associations. Multiple imputation attempts to preserve associations that are present in the data, but it cannot create correlations where none exist. Consequently, there is typically no disadvantage to being liberal when choosing a set of auxiliary variables.

Analysis Phase

The analysis phase involves analyzing the imputed data sets as though they were complete data sets. Analyzing multiple imputed data sets returns multiple parameter estimates (e.g., multiple BMI means). Returning to our earlier example, suppose that the health psychologist is interested in participants' mean BMI after completing the weight loss program. To illustrate the analysis phase, we computed the BMI mean and standard error for each of the four imputed data sets in Table 24.5. This produced the following parameter estimates: $\hat{M}_1 = 33.54$ ($SE_1 = 0.92$), $\hat{M}_2 = 32.75$ ($SE_2 = 0.77$), $\hat{M}_3 = 33.64$ ($SE_3 = 0.94$), and $\hat{M}_4 = 32.86$ ($SE_4 = 0.92$).

Pooling Phase

The pooling phase combines the parameter estimates from the imputed data sets into a single parameter estimate. Rubin (1987) suggests pooling the parameter estimates by taking the mean of the parameter estimates from the imputed data sets. We average the BMI means from the imputed data sets to compute the multiple imputation estimate of the BMI mean:

$$\bar{\theta} = \frac{33.54 + 32.75 + 33.64 + 32.86}{4} = 33.20$$

where $\bar{\theta}$ is the average parameter estimate.

We do not just average the standard errors across the imputed data sets because doing so would not reflect the additional noise that results from the missing data. Rubin (1987) defines the standard error using the within-imputation variance, which quantifies variability within the imputed data sets (i.e., sampling error in complete data sets), and the between-imputation variance, which quantifies variability between the imputed data sets (i.e., additional noise due to the missing scores). The within-imputation variance is the mean of the squared standard errors (i.e., sampling variances) from the imputed data sets:

$$V_W = \frac{1}{m} \sum_{t=1}^m SE_t^2 \quad (24.2)$$

where V_W denotes the within-imputation variance, SE_t^2 denotes the squared standard error from imputed data set t , and m is the number of imputed data sets. The within-imputation variance estimates what the squared standard error would have been had the data been complete.

Analyzing a single imputed data set is problematic because the standard errors do not account for the fact that the imputed scores are just one set of plausible replacement scores. Consequently, standard errors are too small and Type I errors are inflated. By contrast, multiple imputation provides a mechanism for estimating the additional sampling error that results from missing data. The between-imputation variance serves this purpose. The between-imputation variance quantifies the variability of a parameter estimate across the imputed data sets:

$$V_B = \frac{1}{m-1} \sum_{t=1}^m (\hat{\theta}_t - \bar{\theta})^2 \quad (24.3)$$

where V_B denotes the between-imputation variance, $\hat{\theta}_t$ denotes the parameter estimate from imputed data set t , and $\bar{\theta}$ is the pooled (i.e., mean) parameter estimate across the imputed data sets. The sole reason why parameter estimates

vary across imputed data sets is because they contain different imputed scores. Consequently, the between-imputation variance estimates the noise in a parameter estimate attributable to missing data as opposed to random sampling error. Conceptually, you can think of the between-imputation variance as a correction factor that inflates the standard error to adjust for missingness.

Finally, the multiple imputation standard error combines the within-and between-imputation variances, as follows:

$$SE = \sqrt{V_W + V_B + \frac{V_B}{m}} \quad (24.4)$$

where SE denotes the standard error, V_W denotes the within-imputation variance, and V_B denotes the between-imputation variance. Again, the two terms involving the between-imputation variance serve as correction factors that increase the standard error to reflect the additional noise that results from missing data.

Returning to our earlier example, we compute the within-imputation variance of the BMI mean by averaging the squared standard errors from the imputed data

$$V_W = \frac{1}{m} \sum_{t=1}^m SE_t^2 = \frac{0.92^2 + 0.77^2 + 0.94^2 + 0.92^2}{4}$$

sets: $= 0.79$

The within-imputation variance quantifies the random sampling error of the mean. Using the within-imputation variance alone would underestimate the standard error because it reflects random sampling error from the complete data. The between-imputation variance captures the additional noise in the mean estimates that results from missing data. We compute the between-imputation variance by applying the usual formula for the variance to the four mean estimates, as follows.

$$V_B = \frac{1}{m-1} \sum_{t=1}^m (\hat{\theta}_t - \bar{\theta})^2 = \frac{(33.54 - 33.20)^2 + (32.75 - 33.20)^2 + (33.64 - 33.20)^2 + (32.86 - 33.20)^2}{4-1} = 0.21$$

Finally, we then compute the multiple imputation standard error of the BMI

$$SE = \sqrt{V_W + V_B + \frac{V_B}{m}} = \sqrt{0.79 + 0.21 + \frac{0.21}{4}}$$

mean: $= \sqrt{1.05} = 1.03$

Unlike with single imputation methods (e.g., stochastic regression imputation), the multiple imputation standard error accounts for random sampling error as well as additional noise that results from using imputed scores in lieu of observed scores. Correcting the standard error with the between-imputation variance ensures that the Type I error rate does not exceed the nominal significance level (e.g., $\alpha = .05$). Note that we performed the calculations above to demonstrate how the pooling phase of multiple imputation works, but statistical software packages often automate the pooling phase.

Notice that multiple imputation yields a single set of parameter estimates even though we create and analyze multiple imputed data sets. We interpret and report the parameter estimates (e.g., the BMI mean and standard error) from multiple imputation just like the parameter estimates from a complete data set. Enders (2010, chapter 11) gives recommendations for reporting the results of a multiple imputation analysis along with some example write-ups .

Maximum Likelihood Estimation

Methodologists also currently recommend maximum likelihood estimation because, like multiple imputation, it assumes an MAR mechanism. Recall that in multiple imputation we address missing data and then perform our analyses. In maximum likelihood estimation we address missing data while performing our analyses. Maximum likelihood estimation identifies the parameter estimates that maximize the fit of the observed scores to the parameter estimates. Conceptually, compare this to ordinary least squares (OLS) regression analysis, which estimates the regression coefficients that minimize the sum of squared distances between the observed scores and the predicted scores. Maximum likelihood estimation also minimizes the sum of squared distances between the observed scores and the parameter estimates, but it does so using the formula for the normal distribution.

First let us introduce maximum likelihood estimation with a single variable, Y . For univariate data, the log-likelihood is

$$\log L_i = \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-.5\left(\frac{Y_i - \mu}{\sigma}\right)^2} \right] \quad (24.5)$$

where Y_i denotes an observed score for case i and μ and σ are the parameters of

the normal distribution. Collectively, the terms in brackets represent the equation that defines the shape of the normal distribution. Although the log-likelihood equation is complex, we can explain it conceptually. The term in the exponent of the log-likelihood equation contains a squared z-score, which is referred to as the Mahalanobis distance. Because the other terms in the log-likelihood equation are scaling factors that make the area under the normal curve sum to one, we only need to focus on the Mahalanobis distance to understand the log-likelihood equation. The Mahalanobis distance quantifies the standardized distance between an observed score and the mean (i.e., the fit of the observed score to the parameter estimates). As the squared z-score decreases, the fit of the observed score to the parameter estimates increases; numerically, an improvement in fit corresponds to a higher log-likelihood. Maximum likelihood estimation aims to find the parameter estimates (i.e., μ and σ^2 for univariate data) that maximize the fit to the data (i.e., minimize the squared z-scores).

To demonstrate how the log-likelihood equation works, let us first assume that the BMI mean is 33.07 and the variance is 13.81 (i.e., the μ and σ^2 values in Equation 24.5 are known). Consider the BMIs in the hypothetical complete data set in [Table 24.1](#). After we plug $\mu = 33.07$ and $\sigma^2 = 13.81$ into the log-likelihood equation, the only remaining variable is the observed score Y_i . The first participant has a BMI of 38.1 and the second participant has a BMI of 35.4. Plugging $Y_1 = 38.1$ into the log-likelihood equation yields a log-likelihood of -3.148 , and plugging $Y_2 = 35.4$ into the log-likelihood equation yields a log-likelihood of -2.428 . Notice that the log-likelihoods are negative, which results from taking the natural logarithm of the values in brackets (a mathematical convenience that reduces rounding error). Nevertheless, higher (i.e., less negative) log-likelihoods indicate better fit. The log-likelihood for the second participant (-2.428) is higher (i.e., less negative) than the log-likelihood for the first participant (-3.148) because observing a BMI of 35.4 is more likely than observing a BMI of 38.1 in a population with a mean of 33.07. Because the distance between a BMI of 35.4 and the BMI mean is smaller, a BMI of 35.4 fits the parameter estimates better than a BMI of 38.1. Computing the log-likelihood for all 20 participants and summing them gives the sample log-likelihood, which summarizes the fit of the observed scores to the parameter estimates. For univariate data, the sample log-likelihood is just the sum of Equation 24.5 across the N participants, as follows:

$$\log L = \sum_{i=1}^N \log \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-.5\left(\frac{Y_i - \mu}{\sigma}\right)^2} \right] \quad (24.6)$$

Of course, we usually have to estimate the parameters. When the parameter estimates are unknown, maximum likelihood estimation identifies the parameter estimates (often with an iterative search algorithm) that maximize the fit of the observed scores. The maximum likelihood estimates maximize the sample log-likelihood and thus minimize the sum of squared distances (i.e., squared z-scores) between the observed scores and the mean. To illustrate the estimation procedure, [Table 24.6](#) gives the log-likelihoods and squared z-scores for three different plausible population means, including the sample mean, $M = 33.07$. Notice that the log-likelihood and z-scores change with each parameter estimate. For example, the first participant with a BMI of 38.1 has the highest log-likelihood (i.e., smallest squared z-score, best fit) when compared to a population mean of 33.07 because this parameter estimate is closest to 38.1. Conversely, this participant has the lowest log-likelihood (i.e., largest squared z-score, worst fit) when compared to a population mean of 32.00. Because the individual log-likelihoods change with each parameter estimate, so too does the sample log-likelihood. Conceptually, maximum likelihood estimation repeatedly auditions different parameter estimates and uses the sample log-likelihood to identify the parameter estimates that yield the best fit to the data. Considering only the three population means in [Table 24.6](#), a population mean of 33.07 gives the best fit to the data because it maximizes the sample log-likelihood in the bottom row of the table (and thus minimizes the sum of squared z-scores in the bottom row). Not coincidentally, the arithmetic mean produces the best fit to the data because it is identical to the maximum likelihood mean estimate (with complete data).

Table 24.6. *Log-Likelihoods for Three Different Population BMI Means*

BMI	Population BMI Mean					
	$\mu = 32.00$		$\mu = 33.00$		$\mu = 33.07$	
	$\log L$	z^2	$\log L$	z^2	$\log L$	z^2
38.1	-3.579	2.694	-3.173	1.883	-3.148	1.832
35.4	-2.650	0.837	-2.440	0.417	-2.428	0.393
34.0	-2.377	0.290	-2.268	0.072	-2.263	0.063
36.8	-3.066	1.668	-2.754	1.045	-2.735	1.007
31.3	-2.250	0.035	-2.336	0.209	-2.345	0.227
38.2	-3.623	2.783	-3.210	1.957	-3.184	1.905
35.9	-2.782	1.101	-2.536	0.609	-2.522	0.580
34.7	-2.496	0.528	-2.336	0.209	-2.328	0.192
32.6	-2.245	0.026	-2.238	0.012	-2.240	0.016
29.8	-2.407	0.350	-2.602	0.741	-2.619	0.774
38.2	-3.623	2.783	-3.210	1.957	-3.184	1.905
33.5	-2.313	0.163	-2.241	0.018	-2.238	0.013
31.2	-2.255	0.046	-2.349	0.235	-2.358	0.253
27.0	-3.137	1.810	-3.535	2.606	-3.565	2.667
34.9	-2.536	0.609	-2.362	0.261	-2.353	0.242
34.4	-2.440	0.417	-2.303	0.142	-2.296	0.128
28.4	-2.701	0.938	-2.998	1.532	-3.021	1.579
29.1	-2.536	0.609	-2.782	1.101	-2.802	1.141
25.6	-3.714	2.965	-4.214	3.964	-4.252	4.039
32.3	-2.235	0.007	-2.250	0.035	-2.253	0.043
Sums	-54.964	20.657	-54.139	19.007	-54.135	19.000

Although we introduced maximum likelihood estimation with univariate data, researchers typically perform multivariate analyses. For multivariate data, the Mahalanobis distance (squared z-score) is

$$z^2 = (\mathbf{Y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}) \quad (24.7)$$

where \mathbf{Y}_i denotes a vector of observed scores (i.e., participant i 's observed scores on all the variables), $\boldsymbol{\mu}$ denotes the mean vector, and $\boldsymbol{\Sigma}$ denotes the covariance matrix. The -1 superscript on the covariance matrix tells us to take the inverse of the covariance matrix, which is the matrix analog of division. The T superscript tells us to transpose the vector (i.e., turn it into a row instead of a column) so that we can multiply $(\mathbf{Y}_i - \boldsymbol{\mu})$ by itself. Thus, the multivariate equation for the Mahalanobis distance multiplies $(\mathbf{Y}_i - \boldsymbol{\mu})$ by itself (i.e., to square it) and then “divides” by the covariance matrix, making it a squared z-score like the univariate equation for the Mahalanobis distance. With multivariate data, the Mahalanobis distance quantifies the standardized distance between a set of observed scores (i.e., participant i 's observed scores on all the variables) and the mean vector. Consistent with univariate data, when the parameter estimates are

unknown, maximum likelihood estimation identifies the parameter estimates that maximize the sample log-likelihood and thus minimize the sum of squared distances (i.e., squared z-scores) between the observed scores and the mean vector.

Until now, we have discussed maximum likelihood estimation with complete data. With missing data, maximum likelihood estimation does not remove cases with missing scores, nor does it impute the missing scores prior to estimating the parameters. Rather, maximum likelihood estimation with missing data just changes the squared z-score computation in the log-likelihood equation to include only the observed scores for each case. With missing data, the Mahalanobis distance for participant i is

$$z_i^2 = (\mathbf{Y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \quad (24.8)$$

Notice that we just added an i subscript to the mean vector and the covariance matrix. The i subscript denotes that the mean vector and the covariance matrix vary across participants depending on which missing data pattern each participant follows. To demonstrate how maximum likelihood estimation addresses missing data, consider the MAR data set in [Table 24.1](#). For the last participant, we compute the squared z-score using the self-efficacy score and BMI:

$$\begin{aligned} z_i^2 &= (\mathbf{Y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \left(\begin{bmatrix} 66.0 \\ 32.3 \end{bmatrix} - \begin{bmatrix} \mu_{SE} \\ \mu_{BMI} \end{bmatrix} \right)^T \begin{bmatrix} \sigma_{SE}^2 & \sigma_{SE,BMI} \\ \sigma_{SE,BMI} & \sigma_{BMI}^2 \end{bmatrix}^{-1} \left(\begin{bmatrix} 66.0 \\ 32.2 \end{bmatrix} - \begin{bmatrix} \mu_{SE} \\ \mu_{BMI} \end{bmatrix} \right) \\ &= \left(\begin{bmatrix} 66.0 \\ 32.3 \end{bmatrix} - \begin{bmatrix} 32.00 \\ 33.24 \end{bmatrix} \right)^T \begin{bmatrix} 13.77^2 & -29.67 \\ -29.67 & 3.90^2 \end{bmatrix}^{-1} \left(\begin{bmatrix} 66.0 \\ 32.2 \end{bmatrix} - \begin{bmatrix} 32.00 \\ 33.24 \end{bmatrix} \right) = 8.83 \end{aligned}$$

We plugged in the maximum likelihood estimates to demonstrate the calculations. For the first participant, we compute the squared z-score using just the self-efficacy score because the BMI is missing:

$$\begin{aligned} z_i^2 &= (\mathbf{Y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \\ &= (10.0 - \mu_{SE})^T \frac{1}{\sigma_{SE}^2} (10.0 - \mu_{SE}) \\ &= (10.0 - 32.00)^T \frac{1}{13.77^2} (10.0 - 32.00) = 2.55 \end{aligned}$$

Notice that we do not use parameter estimates involving BMI (i.e., the mean, variance, and covariance with self-efficacy) because the first participant's BMI is missing. Despite this, the estimation procedure with missing data is identical to

the estimation procedure we described earlier in this section for complete data: An iterative search algorithm substitutes different parameter estimates (e.g., the mean vector and the covariance matrix) into the sample log-likelihood equation until it identifies the parameter estimates that minimize the sum of squared z-scores (and thus maximize the sample log-likelihood). Thus, maximum likelihood estimation does not remove cases with missing scores or impute the missing scores to estimate the parameters .

How Including the Incomplete Cases Improves Accuracy

Including the incomplete cases improves the accuracy of the parameter estimates relative to excluding the incomplete cases (i.e., listwise deletion), but *how* it achieves this is less obvious. To demonstrate, consider the MAR data set in [Table 24.1](#). Recall that participants with low self-efficacy to lose weight drop out before completing the weight loss program and thus have missing BMIs. Self-efficacy and BMI are negatively correlated, so the observed self-efficacy scores contain information about the missing BMI scores. Maximum likelihood estimation uses data that deletion methods remove from the data set (e.g., the observed self-efficacy scores for cases with missing BMIs) to improve the accuracy of the parameter estimates. In a bivariate normal distribution with a negative correlation, lower self-efficacy scores are less likely to pair with lower BMIs. Consequently, the estimation procedure increases the BMI mean to adjust for the fact that low self-efficacy scores from the incomplete cases most likely pair with higher BMIs. Thus, the BMI mean from maximum likelihood estimation is much closer to the BMI mean from the hypothetical complete data set.

To demonstrate graphically how maximum likelihood estimation works, consider [Figures 24.6, 24.7, and 24.8](#). [Figure 24.6](#) is the bivariate normal distribution for the population of self-efficacy scores and BMIs from which the hypothetical complete data set in [Table 24.1](#) was drawn. To illustrate, we use the mean vector and the covariance matrix from the hypothetical complete data set to form the bivariate normal distribution. Note that the height of the bivariate normal distribution for a given self-efficacy score and BMI corresponds to the likelihood of observing that set of scores. Consider a participant with a self-efficacy score of 37 and a missing BMI. The bivariate normal distribution suggests that, for a given self-efficacy score, there is a distribution of plausible BMIs (because two participants with a self-efficacy score of 37 can have

different BMIs when the correlation between self-efficacy and BMI is less than 1.0). [Figure 24.7](#) is the slice of the bivariate normal distribution in [Figure 24.6](#) corresponding to a self-efficacy score of 37; this is called the conditional BMI distribution because it conditions on a given self-efficacy score. [Figure 24.8](#) is the rotated view of [Figure 24.7](#) so that we can view the conditional BMI distribution more easily. Notice that 32.25 is at the peak of the conditional BMI distribution in [Figure 24.8](#). Thus, a participant whose self-efficacy score is 37 most likely has a BMI of about 32.25. Although maximum likelihood estimation does not actually impute this score into the data set, it adjusts the parameter estimates involving BMI to account for this missing score having a most likely value of 32.25. The estimation procedure implements a similar adjustment for the remaining incomplete cases.

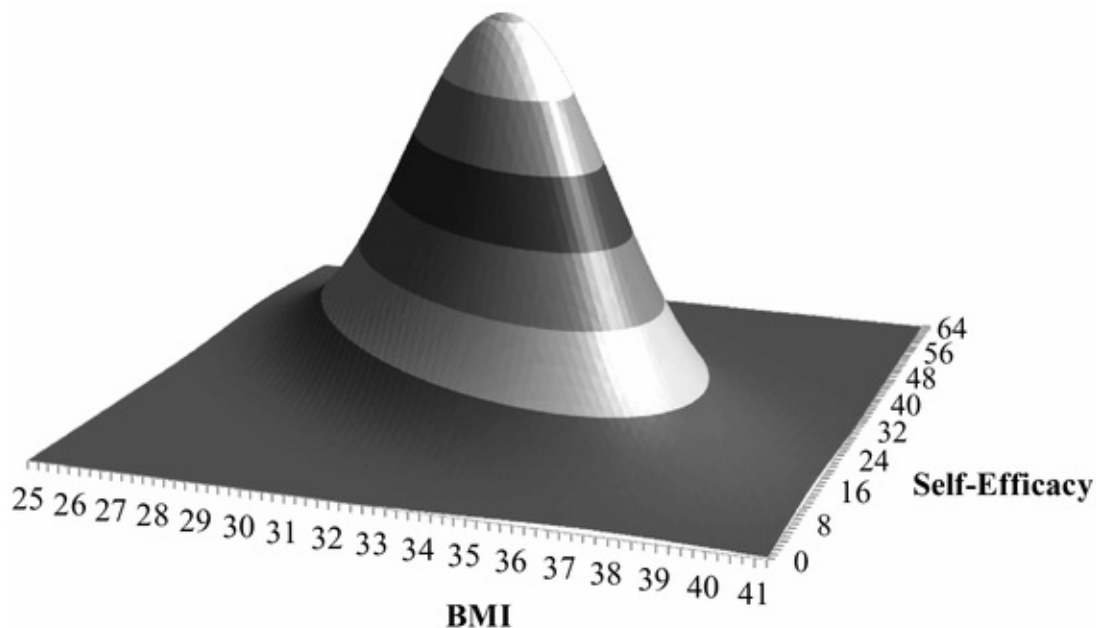


Figure 24.6. Bivariate normal distribution for the self-efficacy scores and BMIs.

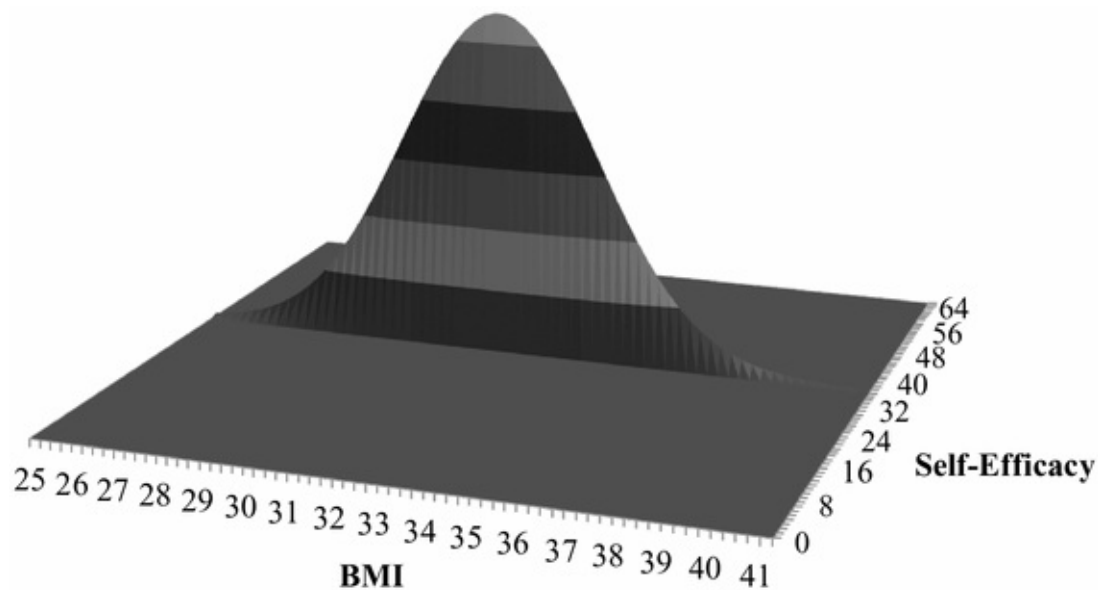


Figure 24.7. Unrotated view of the conditional BMI distribution for a self-efficacy score of 37. The conditional BMI distribution for a self-efficacy score of 37 is just a slice of the bivariate normal distribution in [Figure 24.6](#).

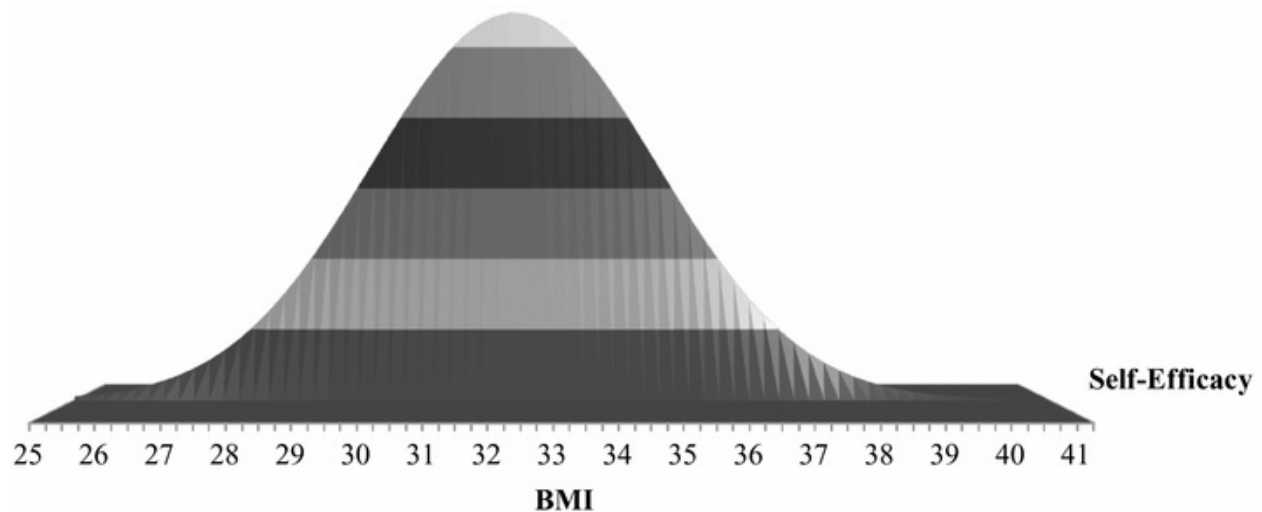


Figure 24.8. Rotated view of the conditional BMI distribution for a self-efficacy score of 37. The conditional BMI distribution for a self-efficacy score of 37 is just a slice of the bivariate normal distribution in [Figure 24.6](#).

Auxiliary Variables

Recall that auxiliary variables predict missingness or correlate with an incomplete variable. As with multiple imputation, methodologists recommend an inclusive analysis strategy that uses auxiliary variables during maximum likelihood estimation. In multiple imputation, we add auxiliary variables as

predictors to the imputation regression equation. In maximum likelihood estimation, we address missing data while performing our analyses; thus, we include the auxiliary variables in our analyses regardless of whether or not we would have included them had the data been complete. To incorporate auxiliary variables into maximum likelihood estimation without changing the interpretation of the parameter estimates, Graham (2003) recommends using the saturated correlates model or the extra dependent variable model. Here, we discuss the saturated correlates model, which is easier to implement. Refer to Graham (2003) for the extra dependent variable model and to Savalei and Bentler (2009) for the two-stage approach for incorporating auxiliary variables (which we also do not discuss here).

For analyses with only manifest variables (i.e., no latent variables), the saturated correlates model correlates an auxiliary variable with (1) predictors, (2) other auxiliary variables, and (3) residuals of the outcome variables. Returning to our earlier example, suppose that the health psychologist estimates a regression model that predicts BMI from self-efficacy. Further, suppose she finds that participants who live farther away from her office drop out before completing the weight loss program and thus have missing BMIs. Finally, suppose that she measured participants' body fat percentages using skinfold measurements, which correlate with BMI. To include distance from the health psychologist's office and body fat percentage as auxiliary variables in a regression equation that predicts participants' BMIs from their self-efficacy scores, we correlate distance from the health psychologist's office and body fat percentage with (1) self-efficacy, (2) each other, and (3) residuals of BMI. Figure 24.9 shows a path diagram of a regression model with two auxiliary variables. Including distance from the health psychologist's office and body fat percentage as auxiliary variables using the saturated correlates model can change the parameter estimates (e.g., by reducing nonresponse bias) but does not change the interpretation of the parameter estimates. For example, the regression coefficient is still the expected change in BMI for a one-unit increase in a participant's self-efficacy score. Graham (2003) gives a similar set of auxiliary variable rules for analyses with latent variables.

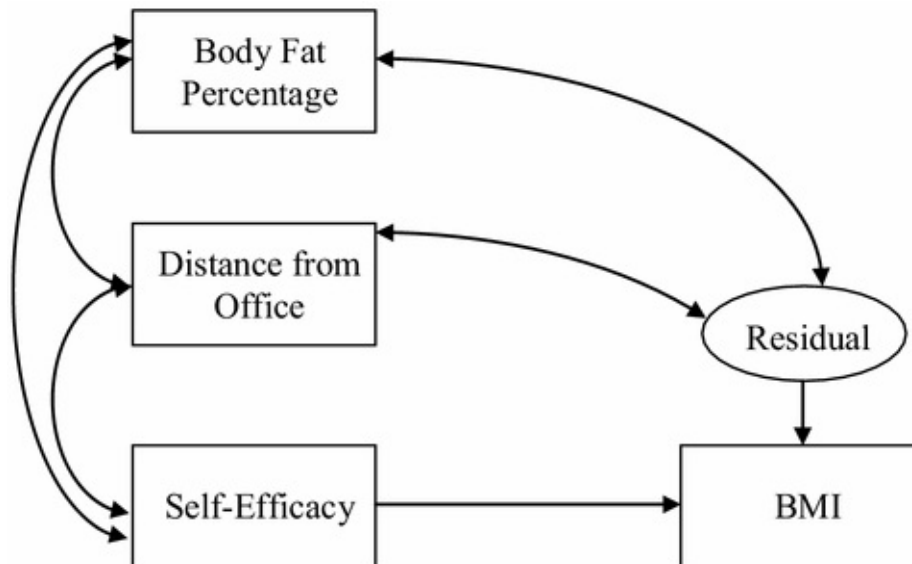


Figure 24.9. Path diagram with auxiliary variables. The straight, single-headed arrow represents a regression coefficient and the curved, double-headed arrows represent correlations.

Comparing Multiple Imputation and Maximum Likelihood Estimation

Unlike most traditional missing data handling methods, multiple imputation and maximum likelihood estimation provide unbiased parameter estimates with an MAR mechanism. Table 24.7 shows that the parameter estimates from these modern missing data handling methods are quite close to those from the hypothetical complete data set. We set up the example to meet the MAR assumption, so the accuracy of the parameter estimates is not surprising and is consistent with missing data theory. However, even when the MAR assumption is violated (meaning the mechanism is NMAR), multiple imputation and maximum likelihood estimation typically provide less-biased parameter estimates than traditional missing data handling methods (Enders, 2010; Schafer & Graham, 2002). Finally, note that the parameter estimates from multiple imputation and maximum likelihood estimation are virtually identical. This is often true, particularly when the maximum likelihood analysis uses the same variables as the imputation phase of multiple imputation. Thus, deciding between these two procedures often depends on practical considerations, which we discuss in the next section .

Table 24.7. *BMI Parameter Estimates from Multiple Imputation and*

Maximum Likelihood Estimation

	Mean	Standard Deviation	Correlation
Complete	33.07	3.72	–.56
Multiple Imputation	33.20	3.87	–.54
Maximum Likelihood Estimation	33.24	3.90	–.55

Practical Considerations

Because multiple imputation and maximum likelihood estimation tend to provide virtually identical parameter estimates, deciding which to use often depends on analysis-specific factors and personal software preferences. These procedures are now available in several general-use statistical software packages, and often they are easy to implement and require relatively short computing times. However, not all statistical software packages offer multiple imputation *and* maximum likelihood estimation. Recall that in multiple imputation we address missing data and then perform our analyses, whereas in maximum likelihood estimation we address missing data while performing our analyses. Multiple imputation is highly compatible with general-use statistical software packages because it does not require a specialized estimation procedure for each analysis; we just use an imputation procedure to fill in the missing scores and then invoke the same procedures used with complete data to analyze the imputed data sets. General-use statistical software packages such as SAS and SPSS generate and analyze multiple imputed data sets and pool the parameter estimates and standard errors.

Because we address missing data while performing our analyses in maximum likelihood estimation, each analysis (e.g., *t*-test, ANOVA, regression analysis) requires a specialized estimation procedure. This makes maximum likelihood estimation less compatible with general-use statistical software packages such as SAS and SPSS (although SAS and SPSS can perform maximum likelihood estimation when computing correlations and means). However, all commercially available structural equation modeling software packages (e.g., Mplus, LISREL,

EQS, and AMOS) perform maximum likelihood estimation. Open-source statistical software packages in R (e.g., Lavaan) also perform maximum likelihood estimation. Many procedures (e.g., ANOVA, regression analysis, correlations) can be recast as structural equation models, so we can use these software packages for analyses that we would typically perform in SAS and SPSS. Some structural equation modeling software packages also perform multiple imputation. To illustrate the process of implementing these techniques, Appendixes A through E at the end of the chapter give Mplus, SAS, and SPSS code for computing descriptive statistics and correlations with maximum likelihood estimation and multiple imputation. Additional syntax examples are available at <http://www.appliedmissingdata.com>.

Choosing Multiple Imputation

Maximum likelihood estimation is arguably easier to implement than multiple imputation, particularly if you are familiar with a structural equation modeling software package. Invoking maximum likelihood estimation often involves adding one keyword or line of syntax, and researchers do not need to understand what goes on in the “black box.” Typically, performing an analysis with missing data in structural equation modeling software packages does not differ from performing an analysis with complete data because maximum likelihood estimation is the default when using raw data (as opposed to summary statistics) as input. Multiple imputation has more procedural steps (e.g., performing diagnostic tests to determine the between-imputation interval, generating imputed data sets, analyzing and pooling the parameter estimates) than maximum likelihood estimation. However, statistical software packages and ease of use aside, multiple imputation is more flexible than maximum likelihood estimation. In what follows we discuss analysis-specific factors that warrant choosing multiple imputation over maximum likelihood estimation despite requiring more procedural steps.

One analysis-specific factor for which multiple imputation is sometimes more flexible than maximum likelihood estimation is incomplete predictors. In the imputation phase, complete variables are predictors and incomplete variables are outcome variables regardless of whether they are predictors or outcome variables in subsequent analyses. Thus, we can perform multiple imputation without regard for a variable's role in the analysis. By contrast, statistical software packages that perform maximum likelihood estimation will sometimes exclude cases with incomplete predictors. This is primarily an issue with multilevel

modeling software packages (e.g., the MIXED procedures in SAS and SPSS). Fortunately, structural equation modeling software packages do not suffer from this limitation and can accommodate missing scores on any variable.

Multiple imputation also handles mixtures of continuous and categorical variables better than maximum likelihood estimation does. Recall that multiple imputation and maximum likelihood estimation assume multivariate normality. However, researchers often use ordinal variables (e.g., Likert items), nominal variables (e.g., ethnicity), and binary variables (e.g., yes/no items), which are not normally distributed. Multiple imputation procedures that can handle mixtures of categorical and continuous variables are available in several general-use statistical software packages (e.g., SPSS, R, Mplus). Although some statistical software packages perform maximum likelihood estimation for incomplete categorical outcome variables under very limited circumstances (e.g., with binary outcome variables), the predominant estimation procedures for categorical outcome variables require an MCAR mechanism.

Multiple imputation is typically more flexible with auxiliary variables than maximum likelihood estimation. Recall that in maximum likelihood estimation we incorporate auxiliary variables using the saturated correlates model, whereas in multiple imputation we add auxiliary variables as predictors in the imputation regression equation. We address the missing data when we generate the imputed data sets and then exclude the auxiliary variables from our analyses, which is more convenient when we have to perform a lot of analyses. Using a large number of auxiliary variables with maximum likelihood estimation also tends to produce convergence problems (i.e., problems estimating the parameters), which is not true for multiple imputation.

Finally, item-level nonresponse usually warrants choosing multiple imputation over maximum likelihood estimation. Researchers routinely analyze scale scores that sum or average items measuring a single construct. We previously discussed why researchers should not average the available items (i.e., use person mean imputation) to compute the scale scores. Multiple imputation is a much better solution. In multiple imputation, we impute the missing items and then compute the scale scores using the imputed scores (Gottschall, West, & Enders, 2012). Because we impute the missing items, we can readily perform analyses that require the item-level responses (e.g., factor analyses and reliability analyses) or scale-level responses (e.g., a regression analysis involving a set of scale scores). Structural equation modeling software packages that implement maximum likelihood estimation can readily perform item-level analyses (e.g., factor

analyses) but are less flexible for scale score analyses unless the researcher is willing to recast the scale score as a latent factor. In our view, this is not an ideal solution because it forces the researcher to change the analysis to accommodate the missing data. When applying multiple imputation to questionnaire data, researchers can perform their analyses using the same procedures they would have used had the data been complete .

NMAR-Based Analyses

Until now, we have discussed traditional missing data handling methods, which assume an MCAR mechanism, and modern missing data handling methods, which assume an MAR mechanism. Unlike the MCAR mechanism, the MAR mechanism is often plausible. However, recall that the data do not provide evidence for or against an MAR or NMAR mechanism because we cannot demonstrate that the probability of missing data on Y is either unrelated (MAR) or related (NMAR) to the values of Y itself after controlling for other variables without knowing what the scores on Y would have been had the data been complete. Violating the MAR assumption in multiple imputation and maximum likelihood estimation produces biased parameter estimates. Thus, methodologists have developed missing data handling methods for an NMAR mechanism, including selection models and pattern mixture models. Although NMAR-based analyses are intuitively appealing, they rely on untestable assumptions that, when violated, can produce parameter estimates that are more biased than parameter estimates from MAR-based analyses (Enders, 2010; Schafer & Graham, 2002). Here, we briefly explain how selection models and pattern mixture models work, although we recommend using them with caution. Refer to Enders (2010, chapter 10) and Enders (2011) for a more comprehensive explanation of NMAR-based analyses.

Recall that the NMAR mechanism states that the probability of missing data on Y is related to the values of Y itself. Notice that the missing data handling methods we have described thus far (particularly multiple imputation and maximum likelihood estimation) address missing data (e.g., by imputation) without explaining or predicting the probability of missing data on Y . This is because the MAR mechanism stipulates that the probability of missing data on Y is already explained by other variables in the analysis. If the mechanism is NMAR, the analysis must also link the values of Y itself to the probability of missing data. Because we cannot directly observe or measure each participant's probability of missing data, we instead use a binary missing data indicator as a

rough proxy (i.e., 1 and 0 for observed and missing scores, as in [Table 24.1](#)). NMAR-based analyses augment the analysis of substantive interest with additional parameters that link the incomplete variable to the binary missing data indicator (or indicators), although selection models and pattern mixture models do this differently.

A selection model combines the analysis of substantive interest with an additional regression equation (or equations) that predicts whether or not participants have missing data (Heckman, [1976](#), [1979](#)). Returning to our earlier example, consider a regression analysis that predicts BMI from self-efficacy. If the health psychologist had reason to believe that the probability of having a missing BMI was related to the values of BMI itself (i.e., an NMAR mechanism), she could implement a selection model that simultaneously estimates a regression equation that predicts BMI from self-efficacy (the analysis of substantive interest) and a logistic regression equation that predicts a binary missing data indicator for BMI (the analysis that predicts the probability of missing data). Because the selection model links the outcome variables from these two regression equations (e.g., BMI and the probability of having a missing BMI), it adjusts for an NMAR mechanism. Although not obvious, the selection model relies on strict, untestable distributional assumptions. Even slight violations of the distributional assumptions produce biased parameter estimates, and other factors can also introduce bias (e.g., collinearity, incorrectly specifying the predictors of the logistic regression equation).

A pattern mixture model links the incomplete variable to the binary missing data indicator (a rough proxy for the probability of missing data) differently. Rather than using the binary missing data indicator as an outcome variable in an additional regression equation, the pattern mixture model stratifies participants into subgroups that share a missing data pattern. Returning to our earlier example, the NMAR data set in [Table 24.1](#) has two missing data patterns: (1) participants with complete data and (2) participants with observed self-efficacy scores and missing BMIs. Next, the pattern mixture model applies the analysis of substantive interest to each subgroup and subsequently computes a weighted average of the subgroup-specific parameter estimates (meaning the number of participants in each subgroup determines the weights used while averaging). Returning to our earlier example, the health psychologist computes the means, variances, and correlation between self-efficacy and BMI for each subgroup and then averages the two sets of parameter estimates, which yields a single set of parameter estimates. Because the parameter estimates incorporate the probability of missing data into the estimation procedure (the pattern mixture model

generates parameter estimates for each missing data pattern), they adjust for an NMAR mechanism. However, like the selection model, the pattern mixture model relies on strict, untestable assumptions. For example, notice that we cannot compute the mean BMI or the correlation between self-efficacy and BMI in one of the subgroups because the BMI scores are completely missing. Consequently, we must specify values for the inestimable parameters or borrow values from another subgroup (e.g., the subgroup with complete data). Perhaps not surprisingly, the validity of the parameter estimates depends on the accuracy of the values specified for the inestimable parameters.

Selection models and pattern mixture models rely on different assumptions (for the selection model, distributional assumptions; for the pattern mixture model, assumptions about the inestimable parameters), so they often produce different parameter estimates. Because their assumptions are untestable, we cannot determine which parameter estimates are more accurate, nor can we determine whether the parameter estimates are more accurate than those from an MAR-based analysis. Because NMAR-based analyses are very sensitive to assumption violations, we typically do not rely on a single NMAR-based analysis. Rather, methodologists recommend applying MAR-based analyses (either multiple imputation or maximum likelihood) and multiple NMAR-based analyses, which is referred to as a sensitivity analysis. If the parameter estimates are similar across missing data handling methods that rely on different assumptions, then we can be more confident in our results .

Summary

Given the pervasiveness of missing data, researchers need to understand the modern missing data handling methods reported in published research articles as well as how to apply these procedures to their research. These procedures are advantageous because they rely on weaker (and more plausible) assumptions than traditional missing data handling methods (i.e., an MAR versus an MCAR mechanism), resulting in more accurate parameter estimates. Multiple imputation and maximum likelihood estimation also increase the statistical power to detect effects because they use all the available data. Not surprisingly, the missing data literature suggests that researchers should no longer rely on traditional missing data handling methods. As multiple imputation and maximum likelihood estimation become increasingly common, editors and reviewers are more likely to expect the use of these procedures in published research articles in the coming years. Thus, we wrote this chapter to better

enable researchers to employ these modern missing data handling methods in their research.

References

- Allison, P. D. (2002). *Missing data*. Newbury Park, CA: Sage.
- Beale, E. M. L., & Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society, Series B*, 8, 27–41.
- Bodner, T. E. (2006). Missing data: Prevalence and reporting practices. *Psychological Reports*, 99, 675–680.
- Brown, R. L. (1994). Efficacy of the indirect approach for estimating structural equation models with missing data: A comparison of five methods. *Structural Equation Modeling: A Multidisciplinary Journal*, 1, 287–316.
- Clark, L., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7, 309–319.
- Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6, 330–351.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
- Dixon, W. J. (1988). *BMDP statistical software*. Los Angeles: University of California Press.
- Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.
- Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, 16(1), 1–16.
- Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 8, 430–457.
- Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research*, 47, 1–25.

- Graham, J. W. (2003). Adding missing data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, 10(1), 80–100.
- Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, 31, 197–218.
- Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8, 206–213.
- Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing data designs in analysis of change. In L. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 335–353). Washington, DC: American Psychological Association.
- Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, 11, 323–343.
- Heckman, J. T. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement*, 5, 475–492.
- Heckman, J. T. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161.
- Jeličić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, 45, 1195–1199.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd Ed.). Hoboken, NJ: Wiley.
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (6th ed.). Los Angeles: Muthén & Muthén.
- Nesselroade, J. R., & Baltes, P. B. (1979). *Longitudinal research in the study of behavior and development*. New York: Academic Press.
- Olinsky, A., Chen, S., Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling.

- European Journal of Operational Research*, 151, 53–79.
- Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74, 525–556.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–592.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473–489.
- Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling*, 16, 477–497.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman & Hall.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33, 545–571.
- van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: Chapman & Hall.
- Widaman, K. F. (2006). Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development*, 71, 42–64.
- Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.
- Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1, 368–376.
- van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: Chapman & Hall.

Appendix A: Example Mplus Syntax for Maximum Likelihood Estimation

```

TITLE:
Maximum Likelihood Estimation for the MAR Data Set in Table 24.1;
DATA:
file = Example Data Set.dat;
VARIABLE:
! efficacy = self-efficacy
! bmicompl = hypothetical complete BMIs
! bmimcar = BMIs with an MCAR mechanism
! bmimar = BMIs with an MAR mechanism
! bminmar = BMIs with an NMAR mechanism
names = efficacy bmicompl bmimcar bmimar bminmar;
! Select the variables used in the analysis.
usevariables = efficacy bmimar;
! Specify the missing value code.
missing = all(-99);
ANALYSIS:
! Specify full maximum likelihood estimation as the missing data
handling method.
estimator = ml;
MODEL:
! Estimate the means.
[efficacy bmimar];
! Estimate the variances.
efficacy bmimar;
! Estimate the covariances.
efficacy with bmimar;
OUTPUT:
! standardized gives the correlation.
standardized (stdyx);

```

Appendix B: Example Mplus Syntax for the Imputation Phase of Multiple Imputation

```

TITLE:
Imputation Phase of Multiple Imputation for the MAR Data Set in
Table 24.1;
DATA:
file = Example Data Set.dat;
VARIABLE:
! efficacy = self-efficacy
! bmicompl = hypothetical complete BMIs
! bmimcar = BMIs with an MCAR mechanism
! bmimar = BMIs with an MAR mechanism

```

```

! bminmar = BMIs with an NMAR mechanism
names = efficacy bmicompl bmimcar bmimar bminmar;
! Select the variables used in the analysis.
usevariables = efficacy bmimar;
! Specify the missing value code.
missing = all(-99);
ANALYSIS:
! Specify the saturated imputation model.
type = basic;
! Specify a random number seed for the MCMC algorithm.
bseed = 57635;
! With a convergence criterion of .05, convergence is achieved when
the
! potential scale reduction (PSR) factor drops below 1.05.
bconvergence = .05;
DATA IMPUTATION:
! Specify the incomplete variable(s) to be imputed.
impute = bmimar;
! Specify the number of imputed data sets.
ndatasets = 4;
! Specify the filename prefix for the imputed data sets.
save = Example Data Set Imputation *.dat;
! Specify the between-imputation interval.
thin = 500;
OUTPUT:
! tech8 gives the potential scale reduction (PSR) factor
convergence diagnostic.
tech8;

```

Appendix C: Example Mplus Syntax for the Analysis Phase of Multiple Imputation

```

TITLE:
Analysis Phase of Multiple Imputation for the MAR Data Set in Table
24.1;
DATA:
! Specify the listing file containing the imputed data set
filenames.
file = Example Data Set Imputation List.dat;
! Imputed data sets are used as input.
type = imputation;
VARIABLE:
! efficacy = self-efficacy
! bmimar = imputed BMIs with an MAR mechanism
names = efficacy bmimar;
! Select the variables used in the analysis.

```

```

usevariables = efficacy bmimar;
ANALYSIS:
! Specify full maximum likelihood estimation as the missing data
handling method.
estimator = ml;
MODEL:
! Estimate the means.
[efficacy bmimar];
! Estimate the variances.
efficacy bmimar;
! Estimate the covariances.
efficacy with bmimar;
OUTPUT:
! standardized gives the correlations.
standardized (stdyx);

```

Appendix D: Example SAS Syntax for Multiple Imputation

```

/* Open the data file.
data example;
infile 'F:\Example Data Set.dat';
input efficacy bmicompl bmimcar bmimar bminmar;
if bmimar = -99 then bmimar = .;
run;

Find the maximum likelihood means and covariance matrix.
proc mi data = example nimpute = 0;
var efficacy bmimar;
em;
run;

Impute the missing scores for the MAR data set. nimpute specifies
the number of imputed data sets. seed specifies a random number
seed for the MCMC algorithm. nbiter specifies the burn-in period
and niter specifies the between-imputation interval.
proc mi data = example nimpute = 4 seed = 57635 out = imputed;
var efficacy bmimar;
mcmc nbiter = 500 niter = 200;
run;

Estimate the correlation and descriptive statistics.
proc corr data = imputed outp = micorrs noprint;
var efficacy bmimar;
by imputation;
run;

Pool the correlation and descriptive statistics across the imputed
data sets.*/
proc sort data = micorrs;
by type name;

```

```

run;
proc means data = micorrs noprint;
var efficacy bmimar;
by type name;
output out = pooledcorrs mean = efficacy bmimar;
run;
proc print data = pooledcorrs;
run;

```

Appendix E: Example SPSS Syntax for Multiple Imputation

Open the data file.

```

get file = 'F:\Example Data Set.sav'.
dataset name example window = front.
dataset activate example.

```

Define the measurement scale. efficacy denotes self-efficacy, bmicompl denotes the hypothetical complete BMIs, bmimcar denotes the BMIs with an MCAR mechanism, bmimar denotes the BMIs with an MAR mechanism, and bminmar denotes the BMIs with an NMAR mechanism.

```

variable level efficacy bmicompl bmimcar bmimar bminmar (scale).

```

Impute the missing BMIs for the MAR data set. Specify the MCMC algorithm (fully conditional specification) using method = fcs. maxiter specifies the between-imputation interval and nimputations specifies the number of imputed data sets.

```

dataset declare imputed.
multiple imputation efficacy bmimar
impute method = fcs maxiter = 200 nimputations = 4
outfile imputations = imputed.
dataset activate imputed.

```

Estimate the correlation and descriptive statistics.

```

correlations
variables = efficacy bmimar
statistics = descriptives
/print = twotail nosig.

```

Chapter twenty-five Mediation and Moderation

Charles M. Judd, Vincent Y. Yzerbyt and Dominique Muller

Our goal in this chapter is to provide an up-to-date and relatively comprehensive treatment of procedures for assessing mediation and moderation in social-personality psychology. Both of these processes enable researchers to ask questions of their data that extend the theoretical scope of inquiry beyond simply establishing some overall experimental effect or some simple relationship between two variables. That is, they both begin to enable researchers to arrive at a more comprehensive theoretical understanding of what produces an effect of interest by probing intricacies of that effect. As such, they are related but distinct analytic tools. They are related in the sense that they permit researchers to probe mechanisms underlying and limiting conditions for effects of interest. And yet, the questions they pose are fundamentally different, in ways that are often confused, and the underlying models are distinct.

Given their importance in developing a theoretical understanding of what produces an effect of interest, it is hardly surprising that the assessment of mediation and moderation is ubiquitous in social and personality psychology. As a result, the literature devoted to procedures for estimating and testing mediation and moderation is vast. While we cover what we consider to be the most important points in this literature, our intention is not to cover this literature exhaustively. Rather, our goal is to discuss basic analytic issues, complexities of interpretation and inference, and underlying assumptions and common pitfalls. Additionally, we provide citations to more in-depth and comprehensive treatments throughout.

The chapter is organized into four main sections. In the first short section we provide basic definitions of both mediation and moderation and illustrate the sort of theoretical questions that their assessment permits the researcher to address. Our emphasis here is on the theoretical definitions that underlie both mediation and moderation, rather than on the technical details of estimation and statistical inference. The second section of the chapter is devoted to a more in-depth treatment of mediation, including underlying assumptions, estimation, statistical inference, and power considerations. Coverage here includes both basic models

assuming homogenous errors and more complex multilevel models that allow grouping and nonindependence of observations. The third section is devoted to a more in-depth treatment of moderation, again including underlying assumptions, estimation, statistical inference, and power considerations. And here too we discuss moderation analyses in the multilevel context, with nested nonindependent observations. In the final section we discuss the integration of these two processes, framed as either moderated mediation or mediated moderation. Again we discuss estimation issues for such models and theoretical interpretations and insights that they permit.

Defining Mediation and Moderation

In order to define mediation and moderation, we start with the presumption that research has established some relationship or effect of theoretical interest. For instance, a social psychologist may have conducted research to demonstrate that social projection – that is, people's tendency to consider that others have the same traits or show the same preferences as oneself – depends on others' group membership. Or a personality researcher may have explored ways in which a particular individual difference – say, extraversion – is related to the tendency to assume leadership roles in small-group settings.

Seldom, however, are researchers content with simply the demonstration of such a relationship or effect. To build a theoretical understanding of social behavior and individual differences more broadly, one must probe the mechanisms that underlie an effect and the limiting conditions for its occurrence. Understanding the mechanisms produces more refined assessments of what the effect really is and how it is produced. Understanding its limiting conditions informs the researcher about necessary and sufficient conditions for its occurrence. These two sorts of understandings – one of mechanisms and one of limiting conditions – are the concerns of mediation analyses and moderation analyses, respectively. That is, the goal of mediation assessment is to explore the underlying mechanisms responsible for an effect of interest, whereas the goal of moderation assessment is to explore the ways in which the magnitude of an effect of interest may depend on other variables.

While the questions addressed via the assessment of mediation and moderation are distinct, it is nevertheless the case that gaining knowledge of mechanisms and limiting conditions extends in similar ways one's theoretical understanding of an effect. If one really understands the mechanisms that

produce an effect, then surely one gains insights into the necessary conditions to produce that effect. That is, if one understands the mechanisms, then it seems likely that one could turn off the effect by inhibiting those mechanisms. And if one really understands the conditions under which an effect is or is not produced, then surely one has gained some insight into the mechanisms responsible for an effect. So a full theoretical understanding of an effect of interest involves both understanding mechanisms (the question of mediation) and understanding limiting conditions (the question of moderation), and the knowledge gained from both of these assessments ultimately must converge.

Because of the fact that the understanding of mechanisms and the understanding of limiting conditions are theoretically intertwined and, in combination, give rise to a full theoretical understanding of the effect of interest, the theoretical questions asked by mediation and moderation procedures can be confusing. However, the analytic procedures for assessing mediation and moderation are different. The former set of procedures examines partial effects controlling for hypothesized mediators. The latter set of procedures examines interactions between the independent variable that produces the effect and some other moderating variable. The distinction in analytic procedures enforces the researcher to think clearly about whether he/she is probing mechanisms or limiting conditions .

Mediation

Basic Analytic Model

Suppose that a researcher wants to study the impact of an independent variable X on a dependent variable Y . Imagine that the independent variable has two levels – a treatment condition and a control condition – and that, in order to permit stronger causal inference, participants have been randomly assigned to one or the other of these conditions. In this context, the total linear effect X on Y is estimated by the slope in the following linear model¹:

$$Y_i = b_{01} + c X_i + e_{1i}$$

The effect of X in this model is represented by the diagram in the top half of [Figure 25.1](#).

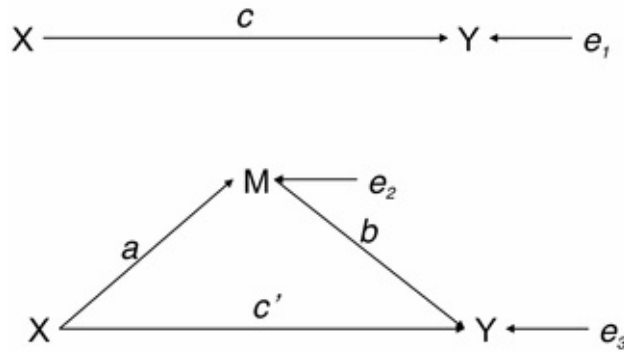


Figure 25.1. Basic mediation model.

In mediational analyses, the researcher is interested in finding the mechanism responsible for this $X - Y$ relationship (Baron & Kenny, 1986; James & Brett, 1984; Judd & Kenny, 1981). Accordingly, the researcher generates hypotheses about one or more third variables that may be partially responsible for the observed total effect, c . The question will then be: Does part of the total effect go through the third variable, often called a mediator or an intervening variable?

To conduct the mediational analysis, one estimates the following two models:

$$M_i = b_{02} + aX_i + e_{2i}$$

$$Y_i = b_{03} + c'X_i + bM_i + e_{3i}$$

In the first of these models, a is the simple effect of X on the mediator. In the second, b is the partial effect of the mediator, controlling for X , and c' is the partial effect of X controlling for the mediator. These models are represented in the diagram at the bottom of Figure 25.1.

The fundamental equation of mediation expresses the total effect c as a function of the coefficients estimated in these two mediational models:²

$$c = ab + c'$$

What this equality tells us is that the total effect of X on Y , c , can be broken into two components, $a*b$ and c' . The first of these components, $a*b$, is the indirect effect of X on Y via the mediator. This is the portion of the total effect that corresponds to the mediation via M . The second of these components, c' , is the residual direct effect of X on Y controlling for or “over and above” the mediator.

It should be noted that the term “direct” must be understood in relative terms,

given that there may be other mediators that potentially explain this residual direct effect (Rucker, Preacher, Tormala, & Petty, 2011). Hence, in the case of two mediators $M1$ and $M2$, the direct effect would be the residual effect of X on Y not explained by either $M1$ or $M2$.

Let us illustrate mediation analysis as well as the underlying models by presenting a concrete example. In a social comparison study, pairs of participants, one of whom was in fact a confederate, performed an attentional task twice (Muller & Butera, 2007, Study 5). After the first round, participants were randomly given bogus feedback: Whereas half of them heard that they had outperformed the confederate (i.e., the downward comparison condition; DC), the other heard they had been outperformed by the confederate (i.e., the upward comparison condition; UC). The self-evaluation threat hypothesis suggests that the UC participants should feel more threatened in their self-evaluation than the DC participants. As a result, they should have fewer attentional resources left in order to process peripheral cues when completing the task. Because peripheral cues could either be selected to help or to hurt participants when dealing with the task, the difference between these two types of cues (called a cuing effect) should be reduced for UC participants. In their study, Muller and Butera did not measure the mediator (i.e., self-evaluation threat) but, in line with their hypotheses, they found a reduced cuing effect among UC than among DC participants.

Imagine now that we conduct a study to examine the hypothesis that self-evaluation mediates the impact of social comparison on participants' attentional resources (the data and SAS codes for this example are available at <http://www.psp.ucl.ac.be/mediation/medmod/>). To this end, we measure self-evaluation right after participants receive the bogus feedback but before they proceed to the second round of the attentional task. As mentioned earlier, the first model allows testing the effect of the independent variable (i.e., social comparison; contrast coded: DC = -0.5 and UC = 0.5) on the dependent variable (i.e., the cuing effect). This analysis reveals a larger cuing effect among DC ($M = 69.33$) than UC participants ($M = 51.65$). Given the coding we use, this translates into a significant negative slope, $c = -17.67$, $t(38) = 2.91$, $p < .01$. In the second model, we test the impact of the independent variable on the mediator (i.e., self-evaluation threat). This analysis reveals a smaller self-evaluation threat in DC ($M = 4.40$) than in UC ($M = 6.32$). Accordingly, this translates into a significant positive slope, $a = 1.92$, $t(38) = 3.62$, $p < .01$. In the last model, we regress the cuing effect on both the independent variable and the mediator. In line with our mediational hypothesis, this analysis reveals a significant slope for

the mediator, $b = -7.88$, $t(37) = 5.80$, $p < .01$, such that (controlling for the independent variable) the higher the self-evaluation threat, the lower the cuing effect. This analysis shows that once we control for the mediator, the effect of the independent variable is no longer significant, $c' = -2.51$, $t(37) = 0.49$, $p = .63$. Finally, in line with the fundamental equation presented earlier (notwithstanding rounding errors), we note that the total effect c equals $a*b + c'$, as $-17.67 = (1.92 * -7.88) + (-2.51)$.

Both the analytical model and the preceding example take for granted that the researcher wants to investigate the mechanism underlying an observed experimental effect. Obviously, the initial step is thus to first establish that such an effect exists. Demonstrating this requires that the experiment have sufficient power to find the overall or total effect. Traditionally, the definition of mediation has taken for granted that there is a significant total effect, and the goal of mediation is then to at least partially account for the process that produces that effect (Baron & Kenny, 1986; Judd & Kenny, 1981).

In recent years, there has been increasing skepticism about the view that the impact of X on Y must be demonstrated before turning to a closer examination of the potential mediating role of a third variable (e.g., Shrout & Bolger, 2002). Relatedly, a similar confusion has surfaced regarding the exact conditions for mediation and whether or not a variable that “suppresses” a total effect should be called a mediator. In the following, we hope to clarify these issues.

One helpful way to think about this is to consider three key features in any situation targeted by a mediational hypothesis. A first feature concerns the presence or absence of a significant c – that is, the total effect of X on Y . A number of reasons may explain why c is not found in the particular data set examined by the researcher. It may be that such an effect simply does not exist. Alternatively, it may exist, but the experiment may not have had sufficient power to detect it. A second feature has to do with c' – that is, the direct residual effect. If in fact the mediator is playing some causal role in affecting Y , then c' should have a different value from c . Finally, the third feature is the indirect effect, $a*b$. When significant, this product points to the existence of a significant causal³ flow between X and Y via the intervening variable, M .

A proper consideration of these three features allows us to define in unambiguous terms what for us corresponds to mediation, suppression, and the mere presence of an indirect effect.⁴ An indirect effect can be said to exist whenever $a*b$ is significant, regardless of the values of c and c' . Both mediation and suppression presume that there is a significant indirect effect, but they imply

additional considerations concerning the magnitudes of c and c' . Mediation for us implies the additional assumption that there is a significant total treatment effect to be explained by the mediational process, i.e., $|c| > 0$, and that this total effect, c , is larger in absolute value than c' . Finally, suppression in the context of a mediational model exists when there is a significant indirect effect and a significant c' , a residual direct effect, that is larger in absolute value than the total effect, c . Suppression means that the intervening variable, when not controlled, is in fact dampening the total effect and that the inclusion of M in the model allows for the direct effect to be more fully revealed.

It should be noted that this discussion focuses on situations in which only one mediator is examined. Matters get somewhat more complex when several mediators are examined. For instance, a third variable may be a suppressor variable and its inclusion in the model could actually reveal the existence not only of a direct effect but also of an indirect effect involving another intervening variable (Rucker et al., [2011](#)).

Assumptions

As with any analysis, mediation analysis with unmanipulated mediators entails a number of important assumptions. A first assumption concerns the requirement that the relations among variables be linear. Of course, nonlinear transformations can be used in the analysis to model nonlinear relations. A second assumption is that the variables are measured both reliably and validly. A third assumption is that the errors or residuals in any one model are independent of each other or, equivalently, that there are no hidden nestings in the data that give rise to dependence. And a fourth and crucial assumption is that the aforementioned models have been correctly specified and that there are no correlated omitted variables that ought to be included in them. This assumption can be equivalently stated as the assumption that the errors or residuals in these models are uncorrelated with the predictor variables included in the models. We will discuss both the second and third assumptions at a later point in the chapter. For now we focus on the fourth assumption because we are convinced that in many applications of mediation analyses it is violated, with serious consequences.

All too often, from our point of view, one finds “mediational” analyses reported using cross-sectional data collected by measuring three variables, X , M , and Y , at roughly the same time. Even assuming no measurement errors, in the absence of any further information, the causal possibilities for why these three variables are related to each other are given by all the straight arrows

(representing potential causal effects) and curved double-headed arrows (representing simple covariances induced by omitted variables) in Figure 25.2. This model obviously includes the possibility that M mediates the $X:Y$ relationship: There is a straight arrow from X to M and another straight arrow from M to Y . But there are also reverse effects that may be responsible for the total covariation observed in the data. For instance, the effect of X on M may result from the impact of X on Y , which in turn affects M . And finally there are also omitted variables that are responsible for the covariation in the errors (the errors are that part of each variable not explained by the direct causal effects to it). If the results of mediational analyses are to provide unbiased estimates of true causal effects, then all of the arrows in Figure 25.2 with the exception of those posited by the mediating process (which we have labeled r , s , and t) must be zero. In other words, a will not equal r unless there is no reverse causal effect of M on X and unless there are no omitted common causes of both M and X . And the same holds for the other effects estimated in a mediational analysis.

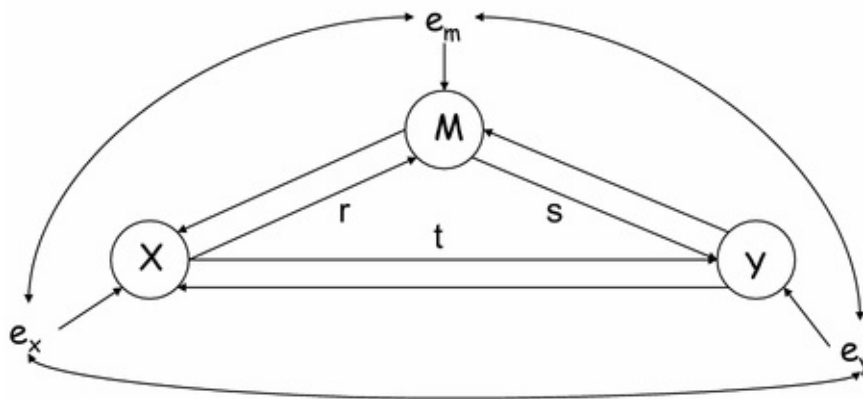


Figure 25.2. Three variable model showing effects responsible for total covariations.

At the beginning of this section we made the assumption that X was an experimentally manipulated independent variable, meaning that participants had been randomly assigned to its levels. The question is now what this buys us in terms of eliminating some of the effects and covariances of Figure 25.2 and thereby improving causal inferences from mediational analyses. With an experimental manipulation of X , the causal possibilities are contained in Figure 25.3. To be sure, many causal possibilities have been eliminated, but it is still the case that there are multiple reasons why the mediator, M , and the outcome variable, Y , covary. And if there is anything other than a direct effect of M on Y , then the mediational estimation will be biased (with bias in both the indirect effect and the residual direct effect).

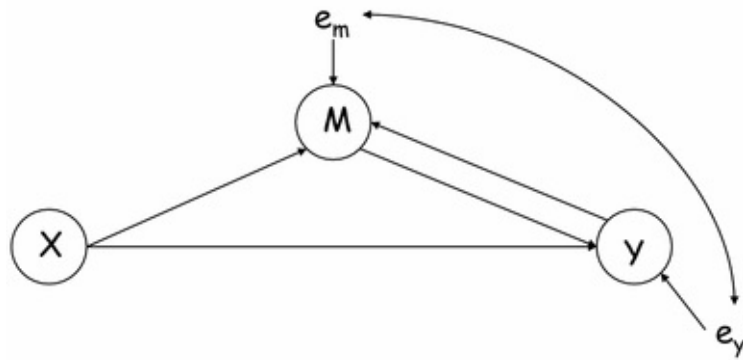


Figure 25.3. Three variable causal model with X manipulated.

What this means is that even with an experimental manipulation of the independent variable, X , mediational analyses will yield biased effects unless there is no potential of Y causing M , and unless there are no omitted common causes of both M and Y . We suspect that in most mediational analyses reported in the literature, even with experimental manipulations of X , the magnitude of the indirect effect via the mediator is substantially overestimated because the mediator and the outcome share omitted common causes. In a great many studies, the outcome and the mediator end up being measured by means of questionnaires, allowing for the intrusion of shared method variance. One of the very early treatments of mediation contained the following warning: The outlined analyses are “likely to yield biased estimates of the causal parameters...even when a randomized experimental research design has been used” (Judd & Kenny, 1981, p. 607, emphasis in the original). Unfortunately, this warning has gone largely unheeded.

MacKinnon (2008) summarizes additional considerations for the single mediator model. Among these, a crucial assumption is temporal precedence, because a mediational model ultimately refers to a causal sequence that must take place across time. As a matter of fact, the mediator model assumes that the treatment variable, X , comes before the mediator, M , which itself comes before the dependent variable, Y . This renders any mediational conclusions based on cross-sectional data highly problematic. The problem more often concerns the ordering of M and Y than the sequence involving X . Related to the issue of temporal precedence are two considerations, namely the level of the mediational chain and the measurement timing. The first concerns the specific steps that are selected for measurement in what may be a rather long and intricate causal chain. Depending on the focus of the researcher, the window used to examine the underlying causal chain may vary widely. The second is related to the

correspondence between the timing of the measurement of the mediator and the dependent variable, on the one hand, and the true timing of the changes in the phenomena under examination, on the other. In many instances, changes in the mediator or in the outcome can occur long after the independent variable has been manipulated.

In Figure 25.4 we include a plausible causal model in the circumstance where X is an experimental manipulation and both M and Y are measured at two time points: time 1 at the same time that X is manipulated and time 2, somewhat later, when the effect of the treatment is thought to have been revealed. Again the mediational indirect effect is the effect of X on $M2$ times the direct effect of $M2$ on $Y2$. Even with such longitudinal data, this indirect effect will be estimated with bias if there is a reverse effect from $Y2$ on $M2$ or if there are omitted third variables responsible for the relationship between $M2$ and $Y2$. This latter threat is reduced in magnitude somewhat because of the fact that earlier values of both variables are controlled, that is, $M1$ and $Y1$. Assuming that other causal effects on these variables are unchanging over time, omitted and unchanging common causes will be effectively controlled by such longitudinal models (in the absence of measurement error) .

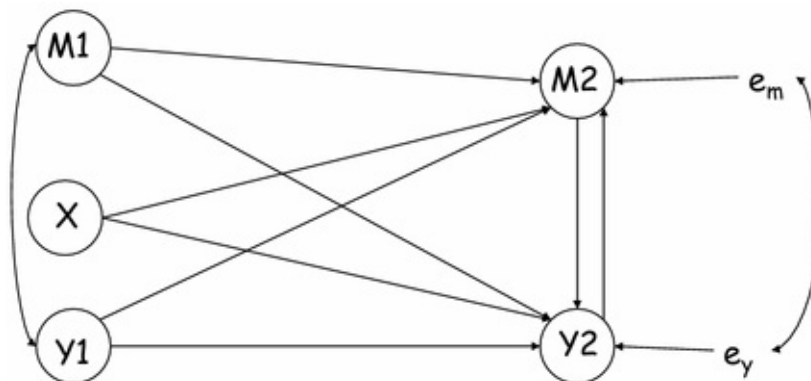


Figure 25.4. Causal possibilities in two-wave longitudinal mediation model.

Estimating and Testing Indirect Effects

So far, we provided the three basic equations underlying mediation analyses. These equations are sufficient to estimate and test the individual slopes in the estimated mediation models (although we later address limitations with these in the presence of measurement error). One important question remains how one should test the underlying indirect effect. There are basically three general approaches: the causal steps, the difference in coefficients, and the product of coefficients (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002).

First, one can test the indirect effect by estimating a and b and testing them individually against zero. The logic here is that if both steps of the indirect effect are significant, it means that the indirect effect is itself significant. In other words, if the two components of a product are significant, the product itself is significant.⁵ Such a test has sometimes been referred to as the a and b joint significance test (Cohen & Cohen, 1983; Fritz, Taylor, & MacKinnon, 2012; MacKinnon et al., 2002).

Second, one can test the indirect effect by estimating and testing the difference between the coefficients c and c' . As we have seen, this test amounts to a test of whether the total X effect is different from the residual direct effect controlling for M . Because of the equivalence of $c - c'$ and $a*b$, this is conceptually similar to asking whether X indirectly influences Y via M .

Third, one can test the indirect effect by directly estimating and testing the product $a*b$. There are three main strategies for doing so. The first consists in testing the $a*b$ product by dividing it by an estimate of the standard error of a product of the two slopes. Baron and Kenny (1986) suggested computing this standard error by using a formula derived by Sobel (1982). The result of this computation is then compared with a normal distribution. As it turns out, this strategy is problematic because the product of two normally distributed variables does not typically have a normal distribution (Bollen & Stine, 1990; MacKinnon et al., 2002). A second strategy makes no assumption regarding the normality of the product distribution and relies on bootstrapping, a resampling method that consists of approximating the population distribution by sampling *with replacement* from the observed sample (Shrout & Bolger, 2002). The simplest bootstrapping technique, the percentile bootstrap, simply uses this bootstrap sampling distribution to provide a 95% confidence interval. The test of the indirect effect is statistically significant if this interval does not include 0. More elaborated bootstrapping techniques – the accelerated bias-corrected bootstrap and the bias-corrected bootstrap – involve corrections for potential biases in $a*b$ estimation and its standard deviation. The third strategy solves the normality issue by using numerical integration to estimate the distribution of this product. MacKinnon and colleagues developed a program called PRODCLIN for doing this (e.g., Fritz et al., 2012; MacKinnon, Fritz, Williams, & Lockwood, 2007). The program enables one to derive a 95% asymmetric confidence interval that can be used to estimate and test $a*b$, again with the indirect effect being statistically significant if 0 is not found in this confidence interval (for more information, consult <http://quantpsy.org/sobel/sobel.htm>).

A fair amount of the mediation literature has been devoted to comparing these indirect effect tests (e.g., Fritz & MacKinnon, 2007; Fritz et al., 2012; MacKinnon et al., 2002). What this work shows is that the most powerful tests are probably the bias-corrected bootstrap tests (e.g., Fritz & MacKinnon, 2007). The major problem, however, is that this increase in power comes with a price in terms of Type 1 error (Cheung, 2009; Fritz et al., 2012; MacKinnon, Lockwood, & Williams, 2004). The increase in Type 1 error, although not dramatic, is especially found when either a or b are 0 and their counterparts are large (i.e., when $a = 0$ and b is large or when a is large and $b = 0$; Fritz et al., 2012). Our own simulations show that the only test that does not suffer from this Type 1 error issue is the a and b joint significance test. This is of interest also because previous work showed that this test is statistically as powerful as the percentile bootstrap and the numerical integration tests (Fritz & MacKinnon, 2007; MacKinnon et al., 2002). The only remaining downside to this approach could be that this test does not provide directly a confidence interval for the indirect effect (MacKinnon et al., 2002). Although having such a confidence interval should not be seen as mandatory, we believe that when necessary one can still estimate indirect effect confidence intervals from either the percentile bootstrap or the numerical integration. Interestingly, MacKinnon and colleagues now suggest that a and b should be tested in addition to using $a*b$ tests (Fritz et al., 2012). Obviously, one additional benefit of the a and b joint significance test, compared to all the other indirect effect tests, is that one can be reasonably confident that each step of the causal path really is significant. As a matter of fact, some data sets may lead to a significant $a*b$ test while either the test of a or the test of b fails to reach significance, calling for some caution in interpretation and for possible replication.

To illustrate the test of the indirect effect, we can go back to the example inspired by the Muller and Butera's (2007) study. If, as we suggest, one relies on the joint significant test, the data reveal that both a and b were significant. Clearly, this test ensures that both a and b are reliable effects. In order to conduct the $a*b$ test, we rely on Preacher and Hayes's (2008) macro. Specifically, the percentile bootstrap (using 5,000 resamples), for instance, gives a $CI_{95\%}$ of -27.94 to -5.98 . Because this confidence interval does not include 0, we can safely conclude that the indirect effect is significant.

Observed versus Latent Variable Models

More often than not, psychologists do not have direct access to their theoretical

variables, but must rely on observed variables instead (e.g., Sigall & Mills, 1998). These observed variables necessarily contain errors of measurement – that part of an observed variable's variance that is not explained by the underlying theoretical construct. The presence of measurement error, as we mentioned earlier, can lead to substantial bias in the estimate of slopes in the mediational models. To reduce measurement errors, social and personality psychologists often measure their theoretical variables with more than one indicator. Researchers are then faced with two main options given that they want to reduce bias from measurement error by using these multiple indicators. First, they can average the multiple indicators and use this summary score as their proxy, their observed variable, in a regression model: This is the observed variable approach for dealing with measurement error (e.g., Ledgerwood & Shrout, 2011). This approach reduces measurement error but does not completely eliminate it (the degree to which such a composite still contains errors of measurement is given by Cronbach's α ; Judd & McClelland, 1998; Schmitt, 1996). Second, researchers can use structural equation modeling to adjust for measurement errors. These models examine relationships among latent variables, adjusted for measurement errors, to test the theoretical model: This is the latent variable approach to estimating mediation (e.g., Ledgerwood & Shrout, 2011). Whatever the chosen model (observed or latent variables), researchers can then proceed with the mediation analysis of their choice presented earlier (MacKinnon, 2008).

To address the pros and cons of these two approaches for dealing with measurement errors, one needs to distinguish between accuracy and precision. Accuracy has to do with how good the model is with respect to estimating the various parameters without the biasing effects of measurement error. In other words, does the model estimate the effects accurately? Precision has to do with the sensitivity of statistical tests of those effects. In other words, does it detect an effect that is in fact present (i.e., does it have a low Type 2 error rate)?

An accurate model is one that provides parameter estimates that are as close as possible to the true theoretical relationships. Latent variable models are better suited for this purpose because they correct for measurement errors. This leads to more accurate estimates. In general, measurement error attenuates bivariate relationships and latent variable models, with multiple indicators, eliminate this attenuation.

The issues, however, become more complex in the context of a three-variable mediational model. For instance, if X is manipulated (hence there is no

measurement error in X) and M is measured with error, the b path will be underestimated in an observed variable model (Hoyle & Kenny, 1999). But importantly, given that $c' = c - ab$, it also means that c' will generally be overestimated (given that b is attenuated). Note that this is just an illustration because matters can become even more complex when X is also measured (and therefore has measurement error; see Ledgerwood & Shrout, 2011). The point remains that latent variable models are better equipped to estimate the true values of the different paths. In a nutshell, latent variable estimates are more accurate.

From the preceding discussion one might surmise that more accurate estimates should also be more precise. For instance, if the effects are underestimated in observed variable models, one would expect them to be significant less often. Paradoxically, this is not what happens. Indeed, Ledgerwood and Shrout (2011) showed that although latent variable models provide more accurate parameter estimates, they also come with larger standard errors, which translate into less precise tests: one will be more likely to conclude that an effect is not there when in fact it exists in the population. Simulations revealed that this is true for both the b and $a*b$ tests. To quote Ledgerwood and Shrout “the latent (vs. observed) variable approach produced estimates that are more accurate but less powerful, especially as reliability decreases and as effect size increases.” (p. 1182). This leads these authors to suggest (apart from wisely encouraging the investment in more reliable measures) a two-step strategy by which one tests the indirect effect with an observed variable model strategy and later estimates this indirect effect with a latent variable model. We concur with this recommendation, particularly when estimating the size of the indirect effect is crucial.

Multilevel mediation

A recent and very important extension of mediational analysis concerns those cases in which the data are collected at more than one level (Krull & MacKinnon, 2001; MacKinnon, 2008; Schoemann, Rhemtulla, & Little, Chapter 21 in this volume). For instance, individuals may be grouped in some way – in work teams, classes, or other naturally occurring groups. When this is the case, if some of the variables in a mediational model are measured at the level of the individuals within those groups, then the assumption that errors or residuals in the models are independent is likely to be violated. This nonindependence arises because observations within a group are likely to be more similar to each other, on average, than are observations between groups, leading to a positive

intraclass correlation due to group. Such dependence can seriously bias statistical inference procedures in tests of mediational models (Schoemann et al., Chapter 21 in this volume).

It is important to bear in mind that there are many plausible groupings that can give rise to dependence in data. In addition to individuals being clustered into groups, observations may be clustered within people, with multiple observations from each person. Such a situation is quite common, with repeated-measures designs that are frequently used by social psychologists. When there are only a few well-defined levels of independent variables of interest that differentiate these repeated measures, procedures that are variations of analysis of covariance with repeated measures can be used to examine issues of mediation within-participants (Judd, Kenny, & McClelland, 2001). In other situations, with more complex designs and numerous observations taken within persons, such as in the case of everyday experience data sets (Reis, Gable, & Maniaci, Chapter 15 in this volume), more general multilevel models are required.

To illustrate multilevel mediational models, we turn to an example from Pleyers, Corneille, Yzerbyt, and Luminet (2009). These authors were interested in factors responsible for evaluative conditioning. More specifically, they wanted to know whether the availability of cognitive resources (the independent variable) plays a role in the emergence of evaluative conditioning (the dependent variable) via its impact on contingency awareness (the mediating variable). Each participant was exposed to several unfamiliar consumption products (conditioned stimuli) that were consistently paired with one of a series of pictures known to elicit either negative or positive affective responses (unconditioned stimuli). All participants wore headphones during the presentation, over which half of them heard music while the remaining half heard numbers and were made cognitively busy by having to perform an auditory two-back task. Specifically, participants were instructed to press the spacebar as quickly as possible when they heard a number that was identical to a number they had heard “two places before” (for instance, if they heard the number “7” and before that they heard a “3” and before that a “7”). The authors then checked the extent to which participants evaluated each product in line with the valence of its associated picture and whether they were able to correctly associate each product with its specific picture. The independent variable, cognitive resources, thus varied between participants. However, both the mediator (contingency awareness) and the evaluative conditioning outcome (whether the specific product was evaluated congruently with the unconditioned stimulus with which it was paired) varied within participants. In short, this study

examined how a level-2 variable (between participants) influences a level-1 variable via a level-1 mediator (within participants). Such a design is formally referred to as a 2–1–1 mediational design (Krull & MacKinnon, 2001).

A proper analysis of such data first requires ascertaining the impact of the level-2 manipulation on the level-1 outcome. This can be done with the following equations:

$$\text{Level 1: } Y_{ij} = d_{0j} + e_{ij}$$

$$\text{Level 2: } d_{0j} = p_{00} + cX_j + u_{0j}$$

In the first of these models, Y_{ij} is the extent to which the evaluation of the i th conditioned stimulus for the j th participant is congruent with its paired unconditioned stimulus (higher numbers mean a stronger evaluation in line with the affective response elicited by the unconditioned stimulus). The level-1 model is in essence estimated for each participant and, accordingly, the estimated intercept, d_{0j} represents the mean degree of evaluative conditioning for each participant and e_{ij} is the variation in that conditioning from product to product within each participant. In the level-2 model, the intercepts from level 1 (mean conditioning for each participant) are modelled as a function of the level-2 experimental condition to which each participant was randomly assigned, X_j . The intercept in this model, p_{00} , is the mean evaluative conditioning on average across participants and c is the degree to which the magnitude of evaluative conditioning depends on the experimental condition. This is the total or unmediated effect of the treatment. u_{0j} is random variation from participant to participant within experimental condition in the magnitude of evaluative conditioning.

The next step involves looking at the impact of the level-2 manipulation on the level-1 mediator. The relevant equations are the following⁶:

$$\text{Level 1: } M_{ij} = d_{0j} + e_{ij}$$

$$\text{Level 2: } d_{0j} = p_{00} + aX_j + u_{0j}$$

These two models are identical to the previous two except now the dependent variable at level-1 is the degree to which the participant is contingency aware for the individual conditioned stimulus (i.e., can he or she state the valence of the unconditioned stimulus with which it was paired during conditioning?). In the

level-2 model the slope of X_j is a , which, parallel to the earlier terms we used for mediation, represents the effect of the treatment on the mediator.

The final step consists in looking at the joint impact of the level-2 manipulation and of the level-1 mediator on the level-1 outcome. The relevant equations are the following:

$$\text{Level 1: } Y_{ij} = d_{0j} + bM_{ij} + e_{ij}$$

$$\text{Level 2: } d_{0j} = p_{00} + c'X_j + u_{0j}$$

As can be seen, these models include one predictor at the within-subject level, M , and one at the participant level, X . The estimate of the b parameter is assessed at the within-subject level, level 1, because the mediator is assumed to be linked to the product, that is, evaluative conditioning with respect to a given product is expected to emerge only when there is awareness of the contingency between this product and the specific unconditioned picture with which it was paired. In the level-2 model c' estimates the residual direct effect of the treatment on the outcome, over and above the mediator.

Multilevel models have a complex structure, most notably because they include errors at multiple levels. As a consequence, the parameters of the model cannot typically be estimated by means of standard least squares methods used for single-level models. Instead, estimation is typically carried out using restricted maximum likelihood (REML) estimation. Having said this, the logic underlying the test of an indirect effect, whether one chooses to rely on $a*b$ or $c - c'$, remains the same even though these estimators will not be exactly equivalent in multilevel models as they are in single-level models. As noted by MacKinnon (2008), the nonequivalence between the two sides of the equation is not really problematic because the discrepancy is likely to be small and decreases as sample sizes increase (Krull & MacKinnon, 1999). Table 25.1 gives an overview of a multilevel analysis of Pleyers et al.'s (2009) data.

Table 25.1. Multilevel Mediation Analysis of Pleyers et al.'s (2009) Data (the analysis was performed using SAS PROC MIXED; data and SAS code are available at <http://www.psp.ucl.ac.be/mediation/medmod/>)

Prediction of level-2 X on level-1 Y

$$\text{Level 1: } Y_{ij} = d_{0j} + e_{ij}$$

$$\text{Level 2: } d_{0j} = 0.1482 - 0.1134 X_j + u_{0j}$$

$$(0.0423) (0.0423)$$

Prediction of level-2 X on level-1 M

$$\text{Level 1: } M_{ij} = d_{0j} + e_{ij}$$

$$\text{Level 2: } d_{0j} = -0.1242 - 0.5665 X_j + u_{0j}$$

$$(0.0546) (0.0546)$$

Prediction of level-2 X on level-1 M and level-1 Y

$$\text{Level 1: } Y_{ij} = d_{0j} + 0.1736 M_{ij} + e_{ij}$$

$$(0.0458)$$

$$\text{Level 2: } d_{0j} = 0.1697 - 0.0151 X_j + u_{0j}$$

$$(0.0411) (0.0483)$$

As can be seen in [Table 25.1](#), the analysis of the data collected by Pleyers *et al.* (2009) corroborates the fact that $a*b = c - c'$ as $-0.5665 * 0.1736 = -0.1134 - (-0.0151)$, which gives $-0.0983 = -0.0983$. [Table 25.1](#) confirms that all the necessary conditions for mediation are satisfied. As a matter of fact, c as well as a and b are significant. Moreover, the standard error of the product can be shown to be .0276. This means that the confidence interval ranges from $-.0442$ to $-.1524$. The fact that the confidence interval does not include 0 confirms the presence of a significant indirect effect. In other words, manipulating the availability of cognitive resources was able to decrease the amount of evaluative conditioning manifested for the various products, and this took place via the impact of cognitive load on participants' contingency awareness, as indexed by their ability to associate each product with the specific image with which it had been paired.

The multilevel approach is a highly flexible one that can be used in a wide variety of designs (Schoemann et al., Chapter 15 in this volume). So, although analytic strategies have been proposed for the examination of repeated measures design with within-subjects manipulations and within-subjects measures of the mediator and of the dependent variable (Judd et al., 2001), such a 1–1–1 design can best be approached from a multilevel perspective. Coming back to our evaluative conditioning example, such a situation would occur if, for every participant in the study, some random set of products had been presented under conditions of cognitive depletion whereas the remaining products had been seen in the absence of a secondary task. A multilevel analysis is of course desirable because the residuals associated with the responses of a particular individual are likely to violate the assumption of independence.

From Measured to Manipulated Mediators

In recent years, several authors have voiced a series of warnings with respect to the possible dividends deriving from a mediational analysis. Indeed, it would seem that mediational analysis has become so popular that it constitutes a mandatory step for any scientific contribution claiming to shed light on a particular psychological process at work in the context of some phenomenon of interest. But does this research really hold its promise? Unfortunately, the answer is not as positive as one would hope it might be. As we discussed earlier, a key problem derives from the fact that a statistical analysis of a set of correlations is taken to confirm the specific causal model put forth by the researcher.

As was noted by early proponents of mediational analysis (Baron & Kenny, 1986; Judd & Kenny, 1981), as well as in more recent contributions (Fiedler, Schott, & Meiser, 2011; MacKinnon, Krull, & Lockwood, 2000; MacKinnon et al., 2002), if a third variable M is indeed a mediator, a logical implication is that its inclusion in the model will reduce the relation between X and Y . At the same time, however, the finding that controlling for M reduces the relation between X and Y does not in fact imply that M is indeed a mediator. Said otherwise, whether a selected causal variable reflects a real cause or not cannot be determined statistically. To be sure, statistical mediation is a necessary condition if one wants to substantiate the conjecture that some third variable is a true mediator, but researchers ought to realize that it is not a sufficient condition.

Fortunately enough, the fact that no correlation pattern can actually prove whether some third variable is causally implicated in the emergence of an effect

does not leave researchers without ammunition. Next to the measurement-of-mediation strategy, several other options allow one to evaluate a causal model whereby some independent variable is thought to set in motion a psychological process that, in turn, produces a given outcome. One prime candidate is the so-called experimental-causal-chain design (Spencer, Zanna, & Fong, 2005). The idea is actually quite simple: When an experimental manipulation, X , is shown to have an impact on some dependent variable, Y , and a specific psychological process, M , is thought to be at work, the researcher is encouraged to decompose the causal sequence into two pieces and conduct an experimental study on each piece. In essence, after demonstrating the c effect, the goal is to conduct two independent experiments addressing both the a and the b effects.

A nice illustration of this strategy comes from a study by Word, Zanna, and Cooper (1974). Building on earlier work on the so-called Pygmalion effect (Rosenthal & Jacobson, 1968), these authors reasoned that people's stereotypes (X) could create a self-fulfilling prophecy (Y) via their nonverbal behavior (M). They first ascertained the a relation by having white participants interview a black or a white confederate. As predicted, participants proved more distant in their nonverbal behavior when facing a black than a white confederate. In a second experiment, white confederates interviewed white unaware participants. As an experimental manipulation of the mediator, interviewers either adopted the distant nonverbal behavior observed with black interviewees in the first experiment or the less distant behavior encountered with white interviewees in the first experiment. In line with the hypothesis, whites did worse on this interview when they were treated like the blacks had been in Experiment 1.

Of course, the experimental-causal-chain design also has limitations. Obviously, the decomposition mandates that the proposed psychological process should be both easy to measure and easy to manipulate. Perhaps the most difficult issue concerns the equivalence between the process as it is measured and the process as it is manipulated. In some cases, it may be difficult to argue compellingly that the dependent variable in the experiment demonstrating the a relation is the same as the independent variable in the experiment demonstrating the b relation. Additionally, the fact that two experiments are conducted in isolation does not allow a proper determination of the amount of variance in the dependent variable accounted for by the independent variable. Keeping these drawbacks in mind, the experimental-causal-chain design still constitutes a powerful tool to uncover the key role that some intervening variable may play along some presumed causal chain.

As we have already discussed, the problem with measuring the mediator rather than manipulating it leaves open the possibility that there are important omitted common causes that can explain the covariation between the mediator and the outcome. There is, additionally, another potential problem with relying on the measurement of the mediator to establish mediation. In many circumstances, a mediator is difficult to measure or its measurement may alter the causal process, either by eliminating the impact of X on Y or by creating an effect (via, for instance, awareness) where none would be observed in the absence of measurement (Jacoby & Sassenberg, 2011; Spencer et al., 2005). Another strategy builds on the realization that mediation rests on the comparison between a factual state, that is, the influence of X on Y , and a counterfactual state, that is, the relation between X and Y when controlling M . In light of this analysis, a smart way to approach the question consists in creating a design in which one compares two factual states. According to Jacoby and Sassenberg (2011; see also Sigall & Mills, 1998), this can be done by means of the testing process by interaction strategy (TPIS), which boils down to an experimental manipulation of the mediator.

Interestingly, the TPIS approach does not require that an experimental effect of X on Y be observed in the first place. If there is an impact of the independent variable in the so-called standard condition, the TPIS would consist in interrupting the process by means of the moderating variable. An alternative plan is to amplify the effect or even reveal an otherwise masked effect by counteracting some suppressing variable. A classic study by Zanna and Cooper (1974) illustrates both processes. These authors had all participants behave counter to their initial attitudes. The implementation of X was straightforward enough: Whereas some were simply forced to do so, others were led to believe that they enjoyed freedom of choice. The extent of attitude change constituted the Y variable. The manipulation of M rested on three conditions. In the control condition, nothing special took place. In line with dissonance theory, free-choice (but not forced) participants experienced unpleasant arousal and changed their attitudes so that they were better aligned with their behavior. In the interruption condition, participants were initially given a (placebo) pill that they were told would cause arousal. This time, free-choice participants did not change their attitudes at all. In the amplification condition, participants were also given a pill but thought that it would cause relaxation. Now free-choice participants modified their attitudes even more than in the control condition. In sum, manipulating the specific way participants experienced the arousal resulting from their counterattitudinal behavior critically affected attitude change,

demonstrating the mediating role of the arousal along with its interpretation .

Promising as the experimental approach may appear, it is of course not in and of itself a panacea to deal with mediation. Some obstacles remain (Bullock, Green, & Ha, 2010). A first issue concerns the isolation of M – that is, the experimental manipulation of M needs to target M and nothing else. A second limitation has to do with the successful manipulation of M . In other words, manipulating M by means of some variable Z is not necessarily equivalent to changing M by means of a manipulation of X . For instance, it may be difficult to be certain that the same people are affected. Finally, researchers ought to realize that there is a possibility for within-sample variation with respect to the indirect effect. In other words, the average indirect effect is potentially misleading. Interestingly enough, this raises the issue of moderated mediation in which the indirect effect varies in magnitude as a function of some other variable. But before we turn to this topic, we turn to a closer examination of moderation .

Moderation

This section focuses on testing and interpreting moderator effects. We start by defining moderation and discuss its relationship with statistical interactions. We then turn to the basic models used to test for interactions and moderator effects. Following this, a major section is devoted to how one interprets the results of these models and best practices for the presentation of moderator effects. We then turn to issues that complicate the search for moderator and interaction effects, focusing in particular on considerations of statistical power. Finally, we briefly discuss some additional issues and designs in which moderation takes on somewhat different forms.

Definitions and Basic Models

As defined earlier, the question of moderation is the question of whether the impact of some independent variable on the dependent variable varies in magnitude as a function of some third variable. Defined in this way, moderation implicitly assumes a causal model in which the independent variable is a cause of the dependent variable and the magnitude of that causal impact depends on some third variable. As such, moderation is not the same as an interaction between two variables. An interaction is the finding that the simple slope of one predictor variable in a linear model varies as a function of the value of another predictor variable. Interactions can exist in the absence of any causal effects of

either predictor variable on the dependent variable.

As we will show analytically, interactions are symmetric: If the simple slope of X on Y varies as a function of Z , then the simple slope of Z on Y varies as a function of X . But when we speak of moderation, we are saying that the $X - Y$ causal relationship is moderated by some variable Z . Because the causal effect that is moderated goes from X to Y , rather from Z to Y , we cannot say that X moderates the $Z - Y$ causal effect. Thus, moderation implies an interaction, but an interaction is not sufficient to claim moderation.⁷ Moderation is an interaction plus the additional strong assumption of a causal impact of an independent variable on a dependent variable that varies in magnitude. The viability of this strong assumption cannot be assessed or confirmed through any data analytic steps. Rather its plausibility depends on theoretical considerations and design variations that permit relatively strong causal inference (e.g., randomized experiments and certain longitudinal designs).

Analytically, a model that estimates moderation is a linear model that estimates the effect of the interaction of two predictor variables on the dependent variable. Interactions are included in models by including as an additional predictor the product of two other predictor variables that are also included in the model. As made clear by Cohen (1978), for the product to estimate the interaction, the two component predictor variables must be included in the model. Thus, the interaction between X and Z is estimated by the slope for the product predictor, b_3 , in the following linear model:

$$Y_i = b_0 + b_1 X_i + b_2 Z_i + b_3 X_i Z_i + e_i \quad (25.1)$$

A test of the interaction is equivalently conducted by testing the null hypothesis that the slope for the interaction is zero, by testing whether this model has a significantly smaller sum of squared errors than the model that does not include the product predictor, or by testing the partial correlations of the product with the criterion (Y), controlling for the two component variables, X and Z (Aiken & West, 1991; Cohen, 1968; Judd, McClelland, & Ryan, 2009).

The slope for the product predictor in the following model, which does not include the two component variables as additional predictors, does not in general estimate the interaction:

$$Y_i = b_0 + b_1 X_i Z_i + e_i$$

To illustrate why it is that the slope for the product predictor in Equation 25.1 estimates the interaction, we can re-express that model as the “simple” relationship between either of the predictor variables and the outcome. Accordingly, the following re-expression represents the “simple” relationship between X and Y at various levels of Z :

$$Y_i = (b_0 + b_2 Z_i) + (b_1 + b_3 Z_i) X_i + e_i \quad (25.2)$$

We can think of these “simple” relationships as simple linear regression models between Y and X whose intercepts and slopes take on different values at varying values of Z . Thus $(b_0 + b_2 Z_i)$ is the simple intercept of these various models and $(b_1 + b_3 Z_i)$ is the simple slope of X . As always, the intercept tells us the predicted Y value when X equals zero, and these simple intercepts vary as a function of the values of Z . And the slope tells us the change in predicted values as X increases by one unit, again with the magnitude of these simple slopes varying as a function of the values of Z .

Given this “simple” re-expression and the crucial centering issues that we will discuss in more detail later in this section, it is important to understand the meaning of the individual slopes in the model in the context of this re-expression. b_0 estimates the intercept of the linear $Y:X$ simple relationship when Z equals zero. b_1 estimates the slope of the linear $Y:X$ simple relationship when Z equals zero. b_2 estimates the change in the intercept of the linear $Y:X$ simple relationship as Z increases by one unit. b_3 estimates the change in the slope of the linear $Y:X$ simple relationship as Z increases by one unit. It is this last parameter estimate that captures the X by Z interaction: To what extent does the simple slope between Y and X change as Z changes in value?

As outlined earlier, interactions are necessarily symmetric, although moderation is not. To demonstrate this symmetry, the model in Equation 25.1 can be equivalently re-expressed as the “simple” relationship between Y and Z :

$$Y_i = (b_0 + b_1 X_i) + (b_2 + b_3 X_i) Z_i + e_i \quad (25.3)$$

with symmetrically equivalent interpretations of the parameter estimates: b_0 estimates the intercept of the linear $Y:Z$ simple relationship when X equals zero. b_2 estimates the slope of the linear $Y:Z$ simple relationship when X equals zero.

b_1 estimates the change in the intercept of the linear $Y:Z$ simple relationship as X increases by one unit. b_3 estimates the change in the slope of the linear $Y:Z$ simple relationship as X increases by one unit.

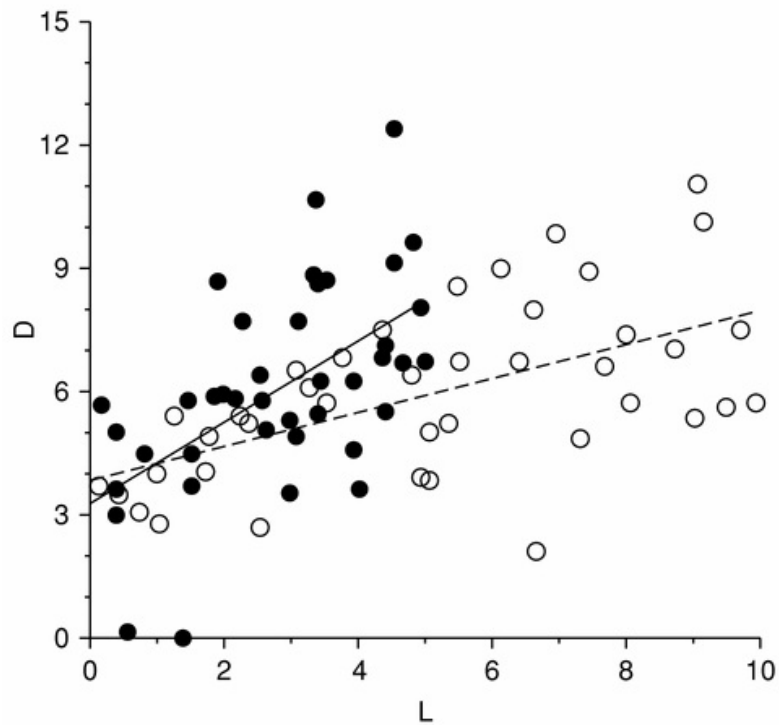
Given that interactions imply that the simple slopes for one predictor take on different values at various levels of another predictor, they imply nonparallel “simple” regression lines. Ordinal interactions are defined as interactions where all the simple slopes for one predictor have the same sign across all meaningful levels of the other predictor. Disordinal or crossover interactions are ones where the simple slopes have both positive and negative values across the meaningful levels of the other predictor.

To this point, we have said nothing about the scale of measurement of either the independent variable or its moderator. When predictor variables are categorical, these need to be coded numerically. Typical coding conventions include dummy coding and contrast coding (also known as effects coding). With two levels of a categorical predictor, the former coding convention uses values of 0 and 1 for the two groups while the latter uses values that sum to zero across the two groups (e.g., $-.5$ and $+.5$). Again the slope of the partialled product of two variables, regardless of whether they are continuously measured or coded categorical variables, will estimate their interaction.

If the moderator variable (Z) is categorical, analyses are frequently reported examining whether the magnitude of the $X:Y$ relationship is different for the different groups, defined by the categorical levels of Z . Frequently this takes the form of testing whether the correlations between X and Y differ across the groups. Such a test is not in general the same as testing whether Z moderates the $X:Y$ relationship, defining moderation as a statistical interaction. Interactions examine whether different simple slopes are needed for the different groups. Correlations in the different groups reflect not only those simple slopes but also the variances of X . It is entirely possible for the different groups to have different simple slopes but the same correlations. And the reverse is entirely possible as well. These two situations are illustrated in the graphs of Figure 25.5 (taken with permission from Whisman & McClelland, 2005). At the top (panel A) we have a situation where the two groups have different slopes but the same correlation. Thus in this case there is an interaction that would be undetected by a comparison of the two group correlations. At the bottom (panel B) there is no interaction – that is, the slopes in the two groups are the same – but one group has a larger correlation than the other does because of relatively greater variance of the X variable. The lesson is that one should test moderation as an interaction

rather than by comparing the magnitude of correlations.

Panel A: Two groups with different slopes (.98 versus .41)
but the same correlations (.57)



Panel B: Two groups with the same slopes (.58)
but different correlations (.38 versus .70)

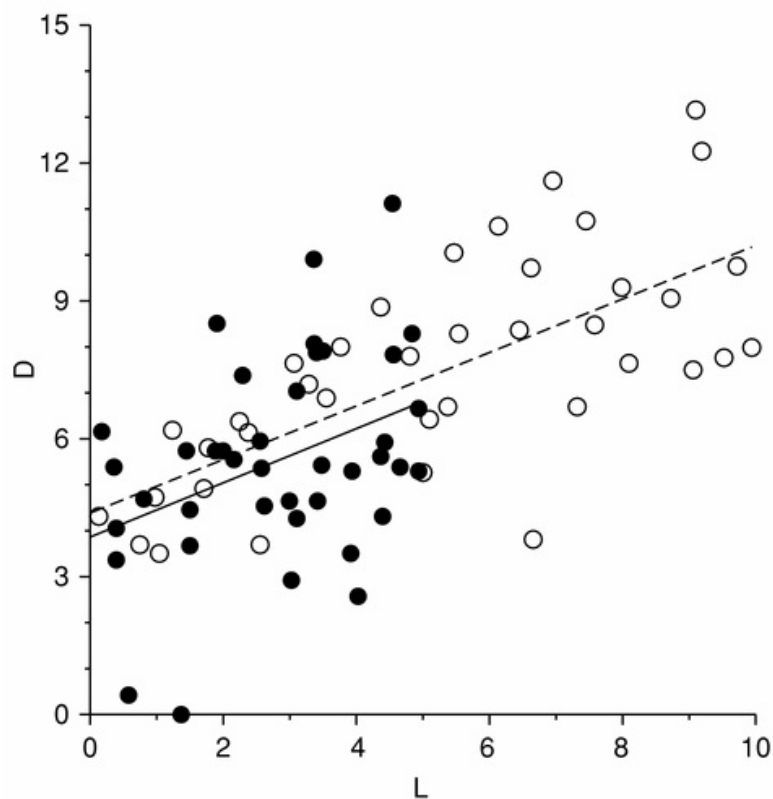


Figure 25.5. Panel A: Two groups with different slopes (.98 versus .41) but the same correlations (.57). Panel B: Two groups with the same slopes (.58) but different correlations (.38 versus .70). Taken with permission from Whisman & McClelland (2005) .

Interpretation and Presentation

Deriving and Plotting Simple Relationships. Once one has found a significant interaction, issues arise as to how that interaction should be interpreted, discussed, and displayed. An initial decision that needs to be made concerns which simple re-expression of the interactive model is theoretically most informative. Recall that there are two such simple re-expressions:

$$Y_i = (b_0 + b_2 Z_i) + (b_1 + b_3 Z_i) X_i + e_i$$

$$Y_i = (b_0 + b_1 X_i) + (b_2 + b_3 X_i) Z_i + e_i$$

If indeed the interaction is because of a moderation of a causal treatment effect – say, for instance, that Z moderates the impact of X on Y – then the choice of the more informative re-expression is easy: The interest is in X and how the simple relationship between X and Y depends on Z (i.e., the first of the preceding two re-expressions). In other cases, where there is not a clear causal model that can be assumed, there is no easy rule to follow in deciding in favor of one or the other of these re-expressions. One should try telling a theoretical story with them both and decide which is the more theoretically interesting and informative. Do you want to argue that the simple relationship between Y and X depends on Z ? Or do you prefer to argue that the simple relationship between Y and Z depends on X ? Both arguments would be correct, but one will generally make a more compelling story than the other.

For now, let us assume that the preferred interpretation is that the simple $Y:X$ relationship depends on Z . Given this, one derives and plots simple $Y:X$ linear relationship predicted by the model at different representative and theoretically meaningful values of Z . If Z is categorical, the choice of these values is easy: One wishes to plot the simple linear relationships for each of the groups defined by the Z categories. If Z is continuously measured, then the choice of the appropriate values of Z for these plots is less clear. One convention derives and plots simple regression lines at the mean of Z and at values of Z one standard deviation above and below the mean (Aiken & West, 1991). For instance,

suppose that values of both X and Z vary between 1 and 5, and the mean of Z is 2.5 with a standard deviation of 1. And suppose the following are the parameter estimates from the interactive model:

$$Y_i = 2.0 + 0.5X_i - 0.5Z_i + 0.3X_iZ_i + e_i$$

The general form of the re-expression of this model is:

$$Y_i = (2.0 - 0.5Z_i) + (0.5 + 0.3Z_i)X_i + e_i$$

At the mean value of Z (i.e., 2.5), the simple $Y:Z$ relationship is:

$$\begin{aligned} Y_i &= (2.0 - 0.5(2.5)) + (0.5 + 0.3(2.5))X_i + e_i \\ &= 0.75 + 1.25X_i + e_i \end{aligned}$$

And at Z values one standard deviation above and below the mean, the simple relationships are:

$$\begin{aligned} Y_i &= (2.0 - 0.5(3.5)) + (0.5 + 0.3(3.5))X_i + e_i \\ &= 0.25 + 1.55X_i + e_i \end{aligned}$$

$$\begin{aligned} Y_i &= (2.0 - 0.5(1.5)) + (0.5 + 0.3(1.5))X_i + e_i \\ &= 1.25 + 0.95X_i + e_i \end{aligned}$$

One then might plot these three simple models, with X on the horizontal axis (with values between 1 and 5) and three different lines, one for each simple relationship. Such a plot is given in [Figure 25.6](#). Inspection of this plot leads to relatively clear interpretations: As X increases, predicted values of Y increase and this is more true at higher levels of Z . That is, the moderation of the $Y:X$ relationship by Z is such that the positive relationship between X and Y becomes larger as Z increases.

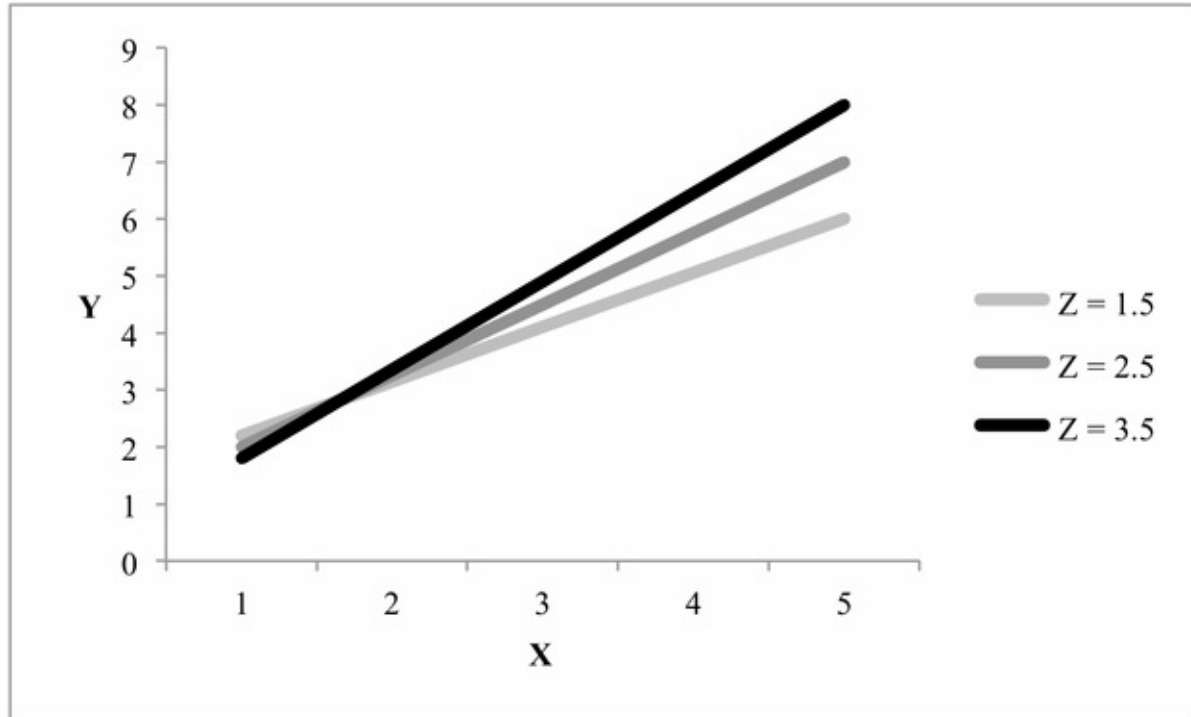


Figure 25.6. Simple Y:X relationships at values of Z of 1.5, 2.5, and 3.5.

There is of course nothing sacred about the choice of Z values at which one chooses to plot these simple relationships. If some theory motivated the desire to see the simple relationship when Z equals 2, one could easily derive and plot this simple relationship in addition to or instead of those at the values of the mean and plus one and minus one standard deviations .

Testing Simple Slopes. A significant interaction tells us that simple slopes vary significantly across the range of values of the moderator variable. But it does not tell us about whether particular simple slopes, at particular values of the moderator variable, are significantly different from zero. Often the further question of whether particular simple slopes are significant is of interest, even though significant simple slopes should not be seen a requirement for interpreting significant interactions (Judd, McClelland, & Culhane, 1995; Rosnow & Rosenthal, 1989).⁸

Given the interactive model and its re-expression in the terms of the Y:X simple relationship:

$$Y_i = b_0 + b_1 X_i + b_2 Z_i + b_3 X_i Z_i + e_i$$

$$Y_i = (b_0 + b_2 Z_i) + (b_1 + b_3 Z_i) X_i + e_i$$

it is apparent that b_1 estimates the simple $Y:X$ slope when Z equals zero. And a statistical test of whether this coefficient differs from zero accordingly provides a test of the simple X slope at that particular value of Z . This then provides a general solution for testing the simple slope of X at any value of Z that might be of interest: One simply deviates Z from that value, recomputes the product term, and tests the X coefficient in the interactive model.

Consider Z' defined as Z deviated from some value c of interest: $Z' = Z - c$. If one now estimated the interactive model using Z' instead of Z (and importantly recomputing the product predictor so that it is XZ'):

$$Y_i = b_0 + b_1 X_i + b_2 Z'_i + b_3 X_i Z'_i + e_i$$

which, re-expressed in terms of the $Y:X$ simple relationship, yields:

$$Y_i = (b_0 + b_2 Z'_i) + (b_1 + b_3 Z'_i) X_i + e_i$$

Accordingly, b_1 now estimates the simple slope of X when Z' equals zero, which it will of course when $Z = c$. In other words, by deviating Z around different values, the slope of X in the model that includes the product predictor takes on different values and these values are the simple X slope at whatever value of c has been used in computing the deviated Z , Z' . And a test that the slope of X differs from zero provides an inferential test of whether the simple slope of X when $Z = c$ differs from zero.

Most commonly, and as recommended by many (e.g., Aiken & West, 1991; Cohen, Cohen, West, & Aiken, 2003; Judd et al., 2009), the value used for c is the mean of Z . This is what is commonly referred to as mean-centering the predictor. Under mean-centering, the slope associated with X will equal the simple slope of X when at the mean of Z in the context of the interactive model that allows the simple slope of X to vary across the values of Z .

Other routinely used values of c include values one standard deviation above and below the mean of Z , as mentioned earlier, thus permitting one to estimate and test simple slopes of X when Z equals those values. But, again, there is nothing sacred about these routinely used values of c . If there are other theoretically meaningful values of Z at which it is important to test whether the simple X slope differs from zero, then the interactive model should be reestimated while deviating Z from those values.

Although the aforementioned approach represents a straightforward procedure

for testing the simple slope of X at different values of Z that are of interest, Preacher, Curran, and Bauer (2006) provide more general procedures that enable the researcher to examine the range of values of Z across which the simple slope of X is and is not significant, given an interactive model. They have implemented their approach in a highly useful Web-based application that is available at <http://quantpsy.org/interact/mlr2.htm> (see also Hayes & Mathes, 2009).

It is worth noting one interesting side consequence of what we have just discussed. We have been considering an additive transformation of one of the variables, $Z' = Z - c$, involved in an estimated interactive model. And what we have shown is that as c takes on different values, the estimated coefficient for X – that is, b_1 – varies, because it generally will equal the simple X slope when Z' equals zero or when $Z = c$. The other estimated slopes in the model, however – b_2 and b_3 – remain constant. Hence, we are left with the curious result that an additive transformation of one of the component predictor variables involved in an interactive model affects the estimated value of the partial slope for the other variable that is a component of the product interaction, but it has no effect on the estimated partial slope of the component variable that is transformed. Additionally, it has no effect on the estimated slope of the product predictor.⁹

Tests of “Main Effects” in Interactive Models. In light of the preceding discussion, it should be clear that the estimated slopes of the component variables (b_1 and b_2) do not in general estimate what the literature devoted to the analysis of variance (ANOVA) refers to as main effects. In the ANOVA literature, a *main effect* represents the effect of one experimental factor on average across the levels of other crossed experimental factors. An estimated slope of a component variable in an interactive model is in general a “simple” slope for that component variable when the other component variable equals zero. If one centers the two component variables on their means (and computes the product predictor as the product of those centered components), then the slope of each component variable will estimate the “simple” slope at the mean value of the other component variable. But even this is not conceptually the same thing as a main effect in the analysis of variance literature. In general, main effects of component variables cannot be defined when those variables are measured more or less continuously and when they are involved in an interaction. An interaction by definition means that there is no “overall” or “main” effect of a component variable. An interaction by definition means that the effects of one component variable vary as a function of the other one.

Accordingly, it is generally a mistake to refer to main effects in moderated regression models. One can certainly estimate and test “simple” effects at the mean value of other component variables, and in many cases these will be very similar to average affects of one variable across the levels of the other variable. But strictly speaking, they are not that; rather they are “simple” effects at the average value of the other variable. Certainly if one fails to center component variables around their means, then slopes associated with those component variables provide nothing resembling what the ANOVA literature defines as main effects.

Standardization. Authors frequently report standardized slopes (or betas) in regression models rather than slopes with the variables in their raw metrics. In general we are not enamored of such practices, for reasons that we briefly discuss but that are reviewed in detail elsewhere (see Turkheimer & Harden, Chapter 8 in this volume; also Cohen, 1990; Tukey, 1969). In simple regression, estimated standardized regression coefficients equal estimated correlations. Hence, as Figure 25.5 illustrated, they are affected by both the magnitude of the raw or unstandardized slope and the variability of the predictor variable. Accordingly, the metrics used in deriving standardized slopes are sample-specific, with the result that standardized slopes will vary from sample to sample, even when raw slopes do not.

In interactive models, these issues are complicated further, as explained by Aiken and West (1991), Freidrich (1982), and Whisman and McClelland (2005). The problem arises from the fact that the product of two standardized variables is not itself standardized (i.e., it will not in general have a mean of 0 and a standard deviation of 1). Accordingly, one might estimate an interactive model and then attempt to interpret the standardized regression coefficient associated with the product predictor. But that slope might be something rather different than if one first standardized the Y , X , and Z variables, computed the product of the two standardized predictors, and then regressed Y on standardized X , standardized Y , and their product. Because of this, we find efforts to interpret standardized slopes in interactive models relatively misleading and at best uninformative. Of course, when a linear regression program outputs those standardized estimates and tests them, those inferential tests are identical to tests of unstandardized or raw regression slopes. But the standardized slopes themselves are interpreted only with difficulty.

Difficulties of Detecting Interactions

The primary challenge in conducting research where moderation is hypothesized is one of assuring adequate statistical power to test interactions. Of course, statistical power is an issue in the conduct of all research (Cohen, 1988), or at the very least should be. Cohen provided guidelines for power consideration when testing “small,” “medium,” and “large” effects. Effect sizes calibrated in this manner are most typically expressed in d units (the ratio of a mean difference to the pooled within group standard deviation). A more general effect size estimate that can be used to calibrate effect sizes is the squared partial correlation, or PRE (Judd et al., 2009), according to which “small,” “medium,” and “large” effects correspond to PREs of .02, .13, and .25 respectively. According to Cohen's power tables, to have adequate power (i.e., $1 - \beta = .80$; $\alpha = .05$), one would need sample sizes of 392, 55, and 26 to detect, respectively, “small,” “medium,” and “large” effects, assuming the absence of measurement error.

This general conclusion about statistical power is complicated in the testing of interactions by two issues that are particularly pernicious in this case. First, unreliability in the measured variables substantially reduces power, and the unreliability of product predictors is a multiplicative function of the unreliability of its component variables (Busemeyer & Jones, 1983; Cohen et al., 2003). Thus, if two variables have reliabilities of .80, the reliability of their product would equal only .64. And as the reliability of a predictor declines, the power needed to test its effect, given some true effect size, is substantially reduced. Aiken and West (1991) estimate that as the reliability of a predictor declines from 1.00 to .80, the sample sizes necessary for acceptable power levels are likely to double. Accordingly, with unreliability of a product predictor being the product of the unreliabilities of its components, sample sizes needed for tests of interactions are likely to be substantially greater than those given earlier.

A second power problem concerns the restriction in range of predictors. In the case of any predictor variable, if it does not vary very much, it is difficult to find an association between it and some outcome variable. This is why in most experimental research, one goes to some length to ensure that the experimental manipulation is substantial enough (i.e., that the difference between a control condition and the treatment condition is large). Outside of the experimental laboratory, it is often difficult to sample respondents who have substantial variability on important predictor variables. Assuming most measured predictor variables have a somewhat unimodal quasi-normal distribution, their variances are typically very substantially less than what the variances would be if everyone

was found only at one or the other extreme value (as we typically construct the distribution to be in experiments).

Thus, whatever factors lead to a restriction of range of a predictor (or, more accurately, a reduction in its variability) decrease the power to find an effect of that predictor. In the case of a product predictor, McClelland and Judd (1993) show that its variance is a multiplicative function of the variance of its components, much like its reliability. Hence factors that restrict or limit the variance of the component variables restrict or limit the variance of the product predictor even more. McClelland and Judd (1993) compare the relative efficiency or power to detect an interaction given various joint distributions of the component variables. In the best-case scenario, all of the observations are located at the most extreme four corners of the joint distribution of X and Z (very high on X and very high on Z , very high on X and very low on Z , etc). This is the ideal design for detecting an interaction, given some constant N , and is the basis for a 2×2 crossed experimental design with manipulations that maximize the variance of the two independent variables. In real-world settings with measured, rather than manipulated, independent variables, it is exceedingly unlikely to encounter such a joint distribution. More likely would be a roughly bivariate normal joint distribution, with most observations clustered near the joint means of the two component variables. McClelland and Judd (1993) showed that such a joint bivariate normal distribution requires 17 times the number of observations to have the equivalent power to detect an interaction compared with the optimal four-corners design.

Thus, when the component variables that are expected to interact are measured variables and when the distribution of each is roughly normal, one will generally have very low power to test an interaction, compared to the optimal design with all observations at the most extreme values of the component variables. As the range of the component variables is reduced, the variance of the product predictor is reduced even more drastically, resulting in a very substantial loss of power. This explains why significant interactions, which are relatively ubiquitous in experimental research, are reported only infrequently using survey or correlational data, unless the sample sizes are exceedingly large (e.g., greater than 500). This then provides some guidance for sampling strategies if one wishes to argue an interactive hypothesis, given measured rather than manipulated independent variables. Rather than sampling randomly, the more powerful alternative is to sample purposively, oversampling the extreme four corners of the joint distribution of the predictors. Some might object that then one moves away from a sample that is truly representative of the

population, which of course is true. But, as always, there are multiple simultaneous and often conflicting goals in research. If one wishes to find significant interactions, then oversampling observations at the extreme four corners of the joint distribution is the most powerful approach.¹⁰

From the preceding discussion readers might draw two erroneous conclusions. The first would be that given a random sample of observations and a hypothesized interaction between two measured predictor variables, one should restrict the analysis to observations that are toward the extreme four corners of the joint distribution. But throwing out observations, regardless of where they are in the joint distribution of the two predictors, will always result in a decrease in statistical power for testing interactions. Given a constant N , it is easier to detect interactions by oversampling the extreme four corners.

A second erroneous conclusion that might be drawn from the fact that interactions are typically found with more power in experiments than with measured predictors is that one should dichotomize predictor variables – using median splits, for instance – and conduct analysis of variance instead of treating measured predictors continuously. There is now a large literature showing that dichotomizing continuous predictors will not result in increases in statistical power (Cohen, 1983; Irwin & McClelland, 2003; MacCallum, Zhang, Preacher, & Rucker, 2002; Maxwell & Delaney, 1993).¹¹

Historically, there have been some who have blamed low power in detecting interactions on multicollinearity between the product predictor variable and its components. In fact, this is not a factor because the collinearity of a product with its component variables is a function of the coding of the component variables, being substantially reduced when the components are centered around their mean. And, as already discussed, such centering has no impact on the test of the interaction in moderated regression models .

Multilevel Interactive Models

In the section of this chapter devoted to mediation, we discussed multilevel models suitable for data where there are nestings of observations that induce dependence (i.e., multiple observations from each participant, observations grouped in classrooms, etc.). Here we provide a short treatment of moderation in such circumstances as well.

As an example, let us turn to a study by Toma, Corneille, and Yzerbyt (2012, Study 3). These authors were interested in social projection and how the

probability of success influences people's tendency to project in cooperative settings. They invited participants into the laboratory and asked them first to self-describe on a series of positive and negative personality traits. Next, participants were told to imagine that they and a partner were involved in a cooperative task allegedly taking place at a software company. Depending on conditions, they learned that the probability of success of the task, and thus of gaining access to a much-desired outcome, was either low (they were informed that, in the past, some 5% of the teams succeeded) or high (95% of the teams succeeded). Finally, participants were presented with the same list of traits as before and asked to describe their partner as well as to indicate the valence associated with each trait.

These data are inherently multilevel, with individual traits that are rated being the first level and the participants being the second. At level 1 – within each participant – there are three variables: participants' ratings of their partner, their rating of themselves, and their rating of the trait's valence. At level 2 – between participants – is the manipulated variable: the probability of success of the cooperative task. Toma *et al.* (2012) expected social projection – that is, that people would generally perceive their partner as being similar to themselves – but that this tendency would be less marked when the probability of success of the cooperation task is thought to be low. The authors also hoped that this pattern would not be affected by the valence of the traits.

For the sake of this presentation, we simplify the analysis somewhat by looking only at social projection for the negative traits (see Table 25.2). At level 1, we consider, that for each one of the j participants, the rating of their partner on the i traits (Y_{ij}) should be predicted by their self-rating on the same traits (X_{ij}).

$$Y_{ij} = b_{0j} + b_{1j}X_{ij} + e_{ij}$$

Table 25.2. Multilevel Moderation Analysis of Toma *et al.*'s (2012, Study 2) Data (the analysis was performed using SAS PROC MIXED; data and SAS code are available at <http://www.psp.ucl.ac.be/mediation/medmod/>)

$$\text{Level 1: } Y_{ij} = b_{0j} + b_{1j}X_{ij} + e_{ij}$$

$$\text{Level 2: } \begin{aligned} b_{0j} &= a_{00} + a_{01}Z_j + u_{0j} \\ b_{1j} &= a_{10} + a_{11}Z_j + u_{1j} \end{aligned}$$

After substitution, we thus have:

$$Y_{ij} = a_{00} + a_{01}Z_j + a_{10}X_{ij} + a_{11}Z_jX_{ij} + u_{0j} + u_{1j}X_{ij} + e_{ij}$$

In the main analysis, this gives:

$$Y_{ij} = 3.836 + 0.046Z_j + 0.293X_{ij} + 0.163Z_jX_{ij} + u_{0j} + u_{1j}X_{ij} + e_{ij}$$

(0.133) (0.133) (0.065) (0.065)

When participants expect success of the cooperation, this becomes:

$$Y_{ij} = 3.882 + 0.046Z_j + 0.456X_{ij} + 0.163Z_jX_{ij} + u_{0j} + u_{1j}X_{ij} + e_{ij}$$

(0.187) (0.133) (0.089) (0.065)

When participants expect success of the cooperation, this gives:

$$Y_{ij} = 3.900 + 0.046Z_j + 0.129X_{ij} + 0.163Z_jX_{ij} + u_{0j} + u_{1j}X_{ij} + e_{ij}$$

(0.188) (0.133) (0.093) (0.065)

As can be seen, each participant has an intercept and a self-rating slope, estimating the impact of the characterization of the self on the characterization of the partner. Greater projection of self-ratings onto partner ratings should be indicated by greater slopes.

At the second level, Toma *et al.* (2012) modeled both the intercepts and the slopes as a function of the probability of success of the cooperative task, contrast-coded (-1, +1) as Z_j :

$$b_{0j} = a_{00} + a_{01}Z_j + u_{0j}$$

$$b_{1j} = a_{10} + a_{11}Z_j + u_{1j}$$

In these level-2 models, the a 's are estimated slopes and intercepts and the u 's are level-2 errors or residuals. The first model is modeling the mean rating¹² of the partner as a function of the probability of success, with a_{01} estimating the degree to which the mean rating of the partner differs between the two experimental conditions. The second of these level-2 models is modeling social projection: to

what extent is the rating of the partner a function of participants' self-rating? The intercept in this second model estimates the mean level of social projection, averaging across participants, and the slope a_{11} estimates the degree to which social projection depends on the experimental manipulation. It is thus this last slope that corresponds to the critical multilevel interaction, that is, the tendency of self-ratings (a level-1 variable) to predict the partner ratings (a level-1 variable) as a function of the manipulated probability of success of the cooperation task (a level-2 variable).

What may seem surprising here is that it is the slope of the level-2 predictor in this second level-2 model that captures the critical interaction, when our expectation up until now has been that slopes of product predictor variables estimate interactions. But it is easy to show that Toma *et al.* (2012) were in fact modeling a product predictor. In the following we have substituted the level-2 estimated models into the level-1 model:

$$\begin{aligned} Y_{ij} &= b_{0j} + b_{1j}X_{ij} + e_{ij} \\ &= (a_{00} + a_{01}Z_j + u_{0j}) + (a_{10} + a_{11}Z_j + u_{1j})X_{ij} + e_{ij} \\ &= a_{00} + a_{01}Z_j + a_{10}X_{ij} + a_{11}Z_jX_{ij} + u_{0j} \\ &\quad + u_{1j}X_{ij} + e_{ij} \end{aligned}$$

Thus, what we ultimately have is a model of social projection as a function of the probability of success, the self-ratings, and the interaction between probability of success and self-ratings. However, this multilevel model differs from the earlier interactive models in that we now have three random-error terms rather than just a single one. First, there is random variance from participant to participant in the mean rating given to the partner (u_{0j}); second, there is random variance from participant to participant in the effect of self-ratings ($u_{1j}X_{ij}$); and finally, there is random variance in individual observations from the participants (e_{ij}). It is the presence of these multiple random-error terms that accommodates the hierarchical nature of the data, allowing for heterogeneity of variance at the different levels.

Table 25.2 presents the output of the PROC SAS analysis of a simplified version of Toma *et al.*'s (2012) data. It can be seen that self-ratings globally predict partner ratings, $a_{10} = 0.293$, $t = 4.53$, $p < .0001$. Importantly, the coefficient associated with the critical interaction term is also significant, $a_{11} = 0.163$, $t = 2.53$, $p = .016$. In line with the authors' predictions, follow-up

analyses, looking at the simple effects of self-ratings on partner ratings in each of the two experimental conditions, confirm that the impact of self-rating proves highly significant when participants expected the cooperation to succeed, $a_{10} = 0.456$, $t = 5.11$, $p < .0001$, whereas there is little trace of social projection when the probability of success was low, $a_{10} = 0.129$, $t = 1.39$, $p = .17$.

In what we have just examined, one of the interacting variables was measured at level 1 and one at level 2. If they are both measured at the same level, either level 1 or level 2, then their interaction would be modeled as a simple product predictor either at level 1 or level 2.

Moderated Mediation and Mediated Moderation

Having now discussed mediation and moderation, we briefly turn to a consideration of models in which both processes are at work. In the case of mediated moderation, Z moderates the overall or total effect of X on Y and the researcher wants to show that some mediator, M , mediates this moderation. The researcher thus wants to show that the moderation is mediated. To illustrate, imagine a persuasion researcher who has shown that attitude change in response to a persuasive communication depends on the interaction of the number of persuasive arguments and their quality: More arguments leads to more persuasion, but only when those arguments are of high quality. It seems reasonable that this interactive effect might be mediated by the depth of processing of the persuasive message. That is, more arguments lead to more in-depth processing, which in turn leads to greater persuasion, but the first link here, from more arguments to more in-depth processing, is mainly found for high-quality arguments.

In the case of moderated mediation, there is an overall treatment effect of X on Y , and the researcher wants to show that the mediation of this treatment effect is different (i.e., moderated) at the different levels of a moderator, Z . The researcher thus wants to show that the mediation is moderated. As an illustration here, imagine a researcher who is interested in the effects of mere exposure on liking. The mediational argument underlying mere-exposure effects might be that more frequent exposure to an object leads to a sense of familiarity, which in turn may lead to greater liking. But the researcher argues that the degree to which this mediational chain may hold should depend on the novelty of the object because the sense of familiarity with the object cannot increase much for objects that are not novel. Thus, the hypothesized mediational path, from

exposure to familiarity to liking, is moderated by novelty.

Interestingly, although the starting questions are different, the basics of mediated moderation and moderated mediation are the same: Both rely on the same underlying models and both imply that the indirect effect (i.e., $a*b$) of the treatment on the outcome via some mediator is moderated by some other Z variable. In other words, the magnitude of $a*b$ depends on Z . Where the two differ, however, is in whether one starts by presuming moderation of the overall or total treatment effect and wants to find a mediator for this moderation or whether one starts by presuming an overall effect and wants to show this overall effect is mediated to a larger extent at different levels of the moderator (Muller, Judd, & Yzerbyt, 2005).

To examine either mediated moderation or moderated mediation, the following models are estimated:

$$\begin{aligned} Y_i &= b_{10} + b_{11}X_i + b_{12}Z_i + b_{13}X_iZ_i + e_{1i} \\ M_i &= b_{20} + b_{21}X_i + b_{22}Z_i + b_{23}X_iZ_i + e_{2i} \\ Y_i &= b_{30} + b_{31}X_i + b_{32}Z_i + b_{33}X_iZ_i + b_{34}M_i \\ &\quad + b_{35}M_iZ_i + e_{3i} \end{aligned}$$

In the following, we assume that in all models Z has been centered on its mean.

In the first of these models, b_{11} estimates the total effect of X on Y at the average level of Z and b_{13} estimates the degree to which that total effect is moderated by Z . In the terminology of mediation, which we gave in the first part of this chapter, b_{11} is equivalent to c , the total effect, allowing that effect to be moderated by Z . Note that in the context of a mediated moderation, although b_{11} is equivalent to c , one is primarily interested in b_{13} and is seeking to explain, via mediated moderation, what the mediating process is that is responsible for the moderation of the overall effect of X on Y .

In the second model, b_{21} estimates the effect of X on the mediator, M , at the average level of Z and b_{23} estimates the degree to which that effect is moderated by Z . In the terminology of mediation given earlier, b_{21} is equivalent to a , the first portion of the indirect effect, again allowing that effect to be moderated.

In the third model, b_{31} is the residual direct effect of the treatment on the outcome at the average level of Z and b_{33} estimates the degree to which that

residual direct effect is moderated. In the earlier terminology, b_{31} is equivalent to c' , allowing this effect to be moderated. Again, note that in the context of mediated moderation, the parameter one is primarily interested in is b_{33} , asking whether the overall moderation of the effect of X , b_{13} , is reduced once one controls for the mediating process (and its moderation).

And finally, also in the third model, b_{34} is the partial effect of the mediator on the outcome controlling for the treatment at the average level of Z and b_{35} estimates the degree to which that effect is moderated. Again, in the earlier terminology b_{34} is equivalent to b , allowing this effect to be moderated.

The resulting mediation models are portrayed in [Figure 25.7](#). The top diagram represents the total effect (the first of the models above), allowing that total effect to be moderated. The bottom diagram represents the second and third models above, allowing all possible effects in this mediational model to be potentially moderated. Earlier, when discussing mediation, we presented the fundamental mediational equality $c - c' = a*b$, with the effects in this equality defined as in [Figure 25.1](#). As shown by Muller *et al.* (2005), there is a similar equality that holds for the mediated moderation and moderated mediation model of [Figure 25.7](#), although now the effect that should be reduced (in the case of mediated moderation) or increased (in the case of a prototypical moderated mediation; see Muller *et al.*, 2005) is not b_{11} (the conceptual analog to c) but b_{13} . Hence, assuming that X is a dichotomous treatment variable that has been contrast-coded (and Z is centered), the equality underlying mediated moderation and moderated mediation is $b_{13} - b_{33} = (b_{23}b_{34}) + (b_{21}b_{35})$. What this equality shows is that the overall moderation of the treatment effect, b_{13} , differs from the moderation of the residual treatment effect on the outcome, b_{33} , as a function of the degree to which the indirect effect is moderated. And in considering whether the indirect effect is moderated, there are two components to consider: whether the effect of the treatment on the mediator is moderated, b_{23} , times the average effect of the mediator on the outcome, b_{34} , and whether the effect of the mediator on the outcome is moderated, b_{35} , times the average effect of the treatment on the mediator, b_{21} .

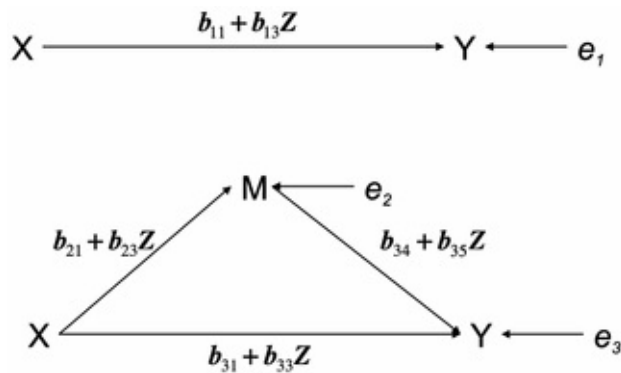


Figure 25.7. Models for mediated moderation and moderated mediation.

Importantly, what this equality further shows is that the moderation of the indirect effect can happen in two ways. First, it may be that the treatment effect on the mediator is moderated (and the mediator affects the outcome). Second, it may be that the mediator's effect on the outcome is moderated (and the treatment affects the mediator). And of course, both of these may be true simultaneously. For us (and others; see also Preacher, Rucker, & Hayes, 2007), this distinction between which component of the indirect effect is moderated is an important theoretical distinction. If there is moderation of the indirect effect via a mediator, then it may be the case that the treatment effect on the mediator is moderated, or it may be the case that the mediator effect on the outcome is moderated.

In sum, because we see this distinction as critical, we suggested that to claim mediated moderation or moderated mediation, in addition to a significant b_{13} (in the case of mediated moderation) or a significant b_{11} (in the case of moderated mediation), researchers need to find either b_{23} and b_{34} , or b_{21} and b_{35} , conjointly significant (Muller *et al.* 2005). Although we do not see it as mandatory (see the Mediation section), one may also want to test whether the overall indirect effect is moderated. To do so, interested readers could refer to the extensive work by Preacher *et al.* (2007), who provide such tests in the context of bootstrapping techniques.

Conclusion

For social and personality psychologists, the techniques for assessing mediation and moderation have become very important tools that are widely used throughout the discipline. Although their use is not without pitfalls, and these have sometimes seriously limited what one can conclude from such analyses, we

are convinced that these are very valuable tools. Their widespread use will continue for the foreseeable future. What we hope to have provided in this chapter is a relatively accessible but thorough guide for the use of these tools and, in so doing, to have clarified underlying assumptions, ongoing controversies, and areas of ambiguity where further work is warranted.

Readers who are familiar with the literature we have reviewed will be aware that some of our definitions, arguments, and suggestions are at variance with definitions, arguments, and suggestions advocated by others whom we highly respect. In our view, this divergence is exciting because it suggests that the last word remains to be written about the wise and appropriate use of mediation and moderation analyses. While these tools are already well developed and widely used, we are convinced they will continue to be refined so that their application will only become more precise and fruitful.

Social and personality psychologists now have at their disposal a wide range of very sophisticated methodological tools that were not in existence some thirty or forty years ago. They should take great pride in these advances. Included in these are analyses to assess mediation and moderation, methods of inquiry that were seldom thought about or practiced only a few decades earlier. Indeed, in these areas, it is social and personality psychologists who have been leading others in the refinement and use of these tools. These are tools of great potential, and their further refinement will continue to be one of our great contributions.

References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, 20, 115–140.
- Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (don't expect an easy answer). *Journal of Personality and Social Psychology*, 98, 550–558.
- Bussemeyer, J. R., & Jones, L. (1983). Analysis of multiplicative combination

rules when the causal variables are measured with error. *Psychological Bulletin*, 93, 549–562.

Cheung, M. W. L. (2009). Comparison of methods for constructing confidence intervals of standardized indirect effects. *Behavior Research Methods*, 41, 425–438.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70, 426–443.

Cohen, J. (1978). Partialled products are interactions; partialled powers are curve components. *Psychological Bulletin*, 85, 858–866.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, 7, 249–253.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 304–312.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression /correlation analyses for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, 47, 1231–1236.

Friedrich, R. J. (1982). In defense of mutliplicative terms in multiple regression equations. *American Journal of Public Health*, 26, 797–833.

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18, 233–239.

Fritz, M. S., Taylor, A. B., & MacKinnon, D. P. (2012). Explanation of two anomalous results in statistical mediation analysis. *Multivariate Behavioral Research*, 47, 61–87.

Hayes, A. F., & Mathes, J. (2009). Computational procedures for probing interactions in OLS and logist regression: SPSS and SAS implementations.

Behavior Research Methods, 41, 924–936.

- Hoyle, R. H., & Kenny, D. A. (1999). Sample size, reliability, and tests of statistical mediation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 195–222). Thousand Oaks, CA: Sage.
- Irwin, J. R., & McClelland, G. H. (2003). Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research*, 40, 366–371.
- Jaccard, J., Turrisi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression*. Newbury Park, CA: Sage.
- Jacoby, J., & Sassenberg, K. (2011). Interactions do not only tell us when, but can also tell us how: Testing process hypotheses by interaction. *European Journal of Social Psychology*, 41, 180–190.
- James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, 69, 307–321.
- Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5, 602–619.
- Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, 6, 115–134.
- Judd, C. M., & McClelland, G. H. (1998). Measurement. In D. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 180–232). New York: McGraw-Hill.
- Judd, C. M., McClelland, G. H., & Culhane, S. E. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. *Annual Review of Psychology*, 46, 433–465.
- Judd, C. M., McClelland, G. H., & Ryan, C. S. (2009). *Data analysis: A model comparison approach* (2nd ed.). New York: Routledge.
- Kraemer, H. C., Kiernan, M., Essex, M., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*, 27, S101–S108.
- Krull, J. L., & MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review*, 23, 418–444.

- Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, 36, 249–277.
- Ledgerwood, A., & Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology*, 101, 1174–1188.
- MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7, 19–40.
- MacKinnon, D. (2008). *Introduction to statistical mediation analyses*. Hillsdale, NJ: Lawrence Erlbaum.
- MacKinnon, D., Fritz, M., Williams, J., & Lockwood, C. (2007). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behavior Research Methods*, 39, 384–389.
- MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, 1, 173–181.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83–104.
- MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39, 99–128.
- Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, 113, 181–190.
- McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114, 376–390.
- Muller, D., & Butera, F. (2007). The focusing effect of self-evaluation threat in coaction and social comparison. *Journal of Personality and Social Psychology*, 93, 194–211.
- Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, 89, 852–863.

- Pleyers, G., Corneille, O., Yzerbyt, V., & Luminet, O. (2009). Evaluative conditioning may incur attentional costs. *Journal of Experimental Psychology: Animal Behavior Processes*, 35, 279–285.
- Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, 31, 437–448.
- Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40, 879–891.
- Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42, 185–227.
- Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston.
- Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin*, 105, 143–146.
- Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass*, 5, 359–371.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8, 350–353.
- Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, 7, 422–445.
- Sigall, H., & Mills, J. (1998). Measures of independent variables and mediators are useful in social psychology experiments: But are they necessary? *Personality and Social Psychology Review*, 2, 218–226.
- Sobel, M.E. (1982). Asymptotic confidence intervals for indirect effects in structural equation modeling. In S. Leinhardt (Ed.), *Sociological methodology 1982* (pp. 291–312). San Francisco: Jossey-Bass.
- Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain:

Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, 89, 845–851.

Toma, C., Corneille, O., & Yzerbyt, V. Y. (2012). Holding a mirror up to the self: Egocentric similarity beliefs underlie social projection in cooperation. *Personality and Social Psychology Bulletin*, 38, 1259–1271.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work. *American Psychologist*, 24, 83–91.

Tzelgov, J., & Henik, A. (1991). Suppression situations in psychological research: Definitions, implications, and applications. *Psychological Bulletin*, 109, 524–536.

Whisman, M. A., & McClelland, G. H. (2005). Designing, testing, and interpreting interactions and moderator effects in family research. *Journal of Family Psychology*, 19, 111–120.

Word, C. O., Zanna, M. P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology*, 10, 109–120.

Zanna, M. P., & Cooper, J. (1974). Dissonance and the pill: An attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology*, 29, 703–709.

¹ Throughout this chapter, for notational simplicity, we express models in terms of parameter estimates rather than the parameters themselves. Estimates may be generated by different estimating procedures in the context of different assumptions about the variables in the models. Most typically they will be least-squares estimates. But, in the case of latent variable models, logistic models involving dichotomous outcomes, or mixed models involving hierarchical levels, estimates will generally be obtained by some maximum likelihood estimation procedure.

² Importantly this equality holds regardless of rescaling of the component variables. Hence, it is also found with standardized estimates.

³ Establishing that this represents a *causal* flow requires a set of very specific assumptions that we detail later in this section.

⁴ In the following, when we talk of suppression, we are doing so within the confines of the causal model underlying mediation, in which some variable M is affected by X and in turn affects Y . Suppression has a broader meaning, i.e., whenever controlling for a third variable augments an effect of interest, outside of the specific causal model that we are assuming (e.g., Tzelgov & Henik, 1991).

⁵ It is possible that some tests of the ab product may yield nonsignificant results even when the two component slopes are significant. This may happen when inappropriate assumptions are made in testing the ab product (i.e., that its sampling distribution is normal).

⁶ Because this mediator is dichotomous rather than continuous, one should be doing estimation of these models using a logistic link function or logistic regression. To keep things simple, however, we chose for this example to act as if the mediator was continuous.

⁷ Our definition of moderation represents a departure from Baron and Kenny's (1986) definition where they equate it with a statistical interaction. We explicitly assume a particular causal model in defining moderation. As such, our definition agrees more closely with recent work devoted to moderation in the context of randomized trials in health outcomes (e.g., Kraemer, Kiernan, Essex, & Kupfer, 2008), where moderation assumes the existence of some causal effect that is moderated.

⁸ The interaction tests whether simple slopes vary, and it is possible that they significantly vary even though particular simple slopes do not differ from zero.

⁹ This final result will not in general be the case if one simply included the product predictor but not its component variable as predictors. It is for this reason that the slope of the product predictor estimates the effect of the interaction only when it is a partialled product, controlling for the two component variables.

¹⁰ Of course, one cannot practically oversample the four corners of the joint distribution if one does not already know respondents' values on those variables. Typically, however, one can identify demographic variables that are likely to covary with the variables, and then the oversampling might be based on those demographics.

¹¹ Also note that researchers sometime dichotomize continuous predictors and argue that it is fine as long as their effects are significant. Because dichotomizing continuous predictors can sometime increase Type-1 errors (Maxwell & Delaney, 1993), we do not recommend this kind of reasoning.

¹² This assumes that the self-ratings have been centered at level 1 around the participant's mean.

Chapter twenty-six MetaAnalysis of Research in Social and Personality Psychology

Blair T. Johnson and Alice H. Eagly*

As in other scientific fields, the progress of social and personality psychology hinges on the orderly and accurate accumulation of empirical evidence about phenomena. This evidence, consisting of multiple studies recording systematic observations of a phenomenon, exists as a literature on the topic. Although new studies rarely replicate earlier studies exactly, many studies are conceptual replications that use different stimulus materials and dependent measures to test the same hypothesis, and still others contain exact replications embedded within designs that add new experimental conditions.

To reach conclusions about empirical support for a particular phenomenon, it is necessary to compare and contrast the findings of relevant studies. Therefore, comparisons of study outcomes – reviews of research – are essential to the scientific enterprise. Until recent decades these comparisons nearly always used methods now known as *narrative reviewing*, which informally draw conclusions about the general trend of the studies' findings, sometimes guided by counts of studies that had either produced or failed to produce statistically significant findings in the hypothesized direction. Such narrative reviews still serve a useful purpose when conducting a comprehensive literature review is not desired or feasible. For example, textbooks typically contain narrative reviews of many hypotheses, and introductions to journal articles reporting primary research usually include brief narrative reviews. These qualitative reviews may suggest useful hypotheses for further scientific investigations.

Despite the usefulness of narrative reviewing, the method does not yield definitive conclusions about the degree of empirical support for a phenomenon or a theory of the phenomenon. One result of this inadequacy is that independent narrative reviews of the same literature often reach differing conclusions. For example, two separate reviews (Brubaker & Powers, 1976; Green, 1981) concluded that those surveyed like younger adults better than older adults, but another review (Lutsky, 1981) concluded that there was little difference. In such cases, it is difficult to determine which conclusion is more accurate.

Critics have pointed to four general faults in narrative reviewing (e.g., Cooper, 2010; Eagly, 1987; Rosenthal, 1991). Although these faults are not necessarily inherent in narrative reviewing, they typify narrative reviewing in practice: (1) Narrative reviewing generally involves the use of a convenience sample of studies, perhaps consisting of only those studies that a reviewer happens to know. Any criteria by which a reviewer selected these studies typically go unstated and may never have been formalized by the reviewer. (2) Narrative reviewers generally do not state their procedures for cataloging studies' characteristics or evaluating the quality of the studies' methods. Any rules or procedures are often not applied uniformly to all of the studies in the sample. (3) When study findings differ, narrative reviewers have difficulty reaching clear conclusions about whether differences in study methods explain differences in results. Because such reviewers usually do not systematically code studies' methods, their procedures are poorly suited to account for inconsistencies in findings. (4) Narrative reviewers typically rely on statistical significance to judge studies' findings and not on the magnitude of the findings. Statistical significance is a poor basis for comparing studies that have different sample sizes because effects of identical magnitude can differ in statistical significance. Because of this problem, narrative reviewers often reach erroneous conclusions about the confirmation of a hypothesis in a series of studies, even in literatures as small as 10 studies (Cooper & Rosenthal, 1980). All four of these problems can render narrative reviews inadequate in most contexts in which research is aggregated and integrated.

These potential flaws in the review process become increasingly aggravated as the number of studies available mounts. In contemporary psychology, large research literatures are not uncommon. For example, even as early as 1978, there were at least 345 studies examining interpersonal expectancy effects (Rosenthal & Rubin, 1978). Similarly, by 1983, there were more than 1,000 studies evaluating whether birth order relates to personality (Ernst & Angst, 1983). As the number of studies increases, the conclusions reached by narrative reviewers typically become more unreliable because of the informality of their methods (Johnson & Boynton, 2008).

Because of the importance of comparing study findings accurately, scholars have dedicated considerable effort to making the review process as reliable and valid as possible and thereby avoiding the criticisms that narrative reviews often engender. The result has been the emergence of review techniques that summarize scientific literatures by methods that are themselves consistent with scientific norms. *Quantitative research synthesis* or *metaanalysis* statistically

cumulates the results of independent empirical tests of a particular relation between variables. More recently, integrative data analysis of individual-level data has also emerged (e.g., Cooper & Pattall, 2009). Although scientists have cumulated empirical data from independent studies since the early 1800s (Stigler, 1986), relatively sophisticated techniques emerged only after the advent of standardized indexes such as r -, d -, and p -values. In the first published monograph related to these strategies, Glass, McGaw, and Smith (1981) emphasized that reviewing scientific literature is a scientific practice that should follow disciplined and transparent steps. Reflecting the maturation of metaanalysis, Hedges and Olkin (1985) presented a sophisticated version of its statistical bases. Standards for metaanalysis have grown increasingly rigorous, as apparent in the two editions of *The Handbook of Research Synthesis and MetaAnalysis* (Cooper & Hedges, 1994; Cooper, Hedges, & Valentine, 2009).

Social psychologists' first rudimentary applications of quantitative review techniques occurred in the 1960s (e.g., Rosenthal, 1968; Wicker, 1969), but it was not until the late 1970s and early 1980s that scholars applied these techniques to a wide range of social psychological phenomena (e.g., Bond & Titus, 1983; Cooper, 1979; Hall, 1978). In many instances, metaanalyses have overturned or enhanced prior narrative reviewers' conclusions. As one example, Sidanius, Pratto, and Bobo (1994) proposed the gender invariance hypothesis – that, across cultures, males score higher in social dominance orientation than do females. Lee, Pratto, and Johnson's (2011) metaanalysis revealed gender differences that varied considerably in magnitude but did not disappear across the cultures investigated. Within social and personality psychology, as in many other sciences, quantitative research synthesis is now well accepted because scholars realize that careful application of these techniques yields the clearest conclusions about a research literature (Card, 2012; Cooper et al., 2009).

To provide a general introduction to metaanalysis, in the remainder of this chapter we (1) present the steps involved in synthesizing research, (2) consider some options that reviewers should consider as they proceed through these steps, (3) discuss standards for conducting and evaluating quantitative reviews, and (4) evaluate metaanalysis relative to primary research and other methods of testing hypotheses. In treating this subject, consistent with convention, we use the term “metaanalysis” to refer broadly to the entirety of the process, including both quantitative and qualitative aspects.¹

Procedures for MetaAnalysis

An Overview of the Process of Quantitative Synthesis

The research process underlying quantitative synthesis can be broken into discrete steps (Cooper, 2010). Each stage contributes to the next stage; careful work in the early stages makes the later stages easier to accomplish and improves the quality of the overall review. As a preview to a more detailed exposition, we list the stages and some of the questions that often accompany them:

1. *Conceptual analysis of the literature.* What independent and dependent variables define the phenomenon? How have these variables been operationalized in research? Have scholars debated different explanations for the relationship demonstrated between these variables? Can the metaanalysis address these competing explanations? When, how much, and in what pattern should the variables relate? Should the size of the relation be relatively consistent or inconsistent across studies?
2. *Setting boundaries for the sample of studies.* What criteria should be used to select studies for the sample? Should considerations of study quality play a major role? What criteria should *exclude* studies from the sample?
3. *Locating relevant studies.* What strategies will best locate the universe of studies? How can unpublished studies be obtained?
4. *Creating the meta-analytic database.* Which study characteristics should be represented, and how can these characteristics be coded or otherwise assessed? How can the quality of a study's methods be assessed?
5. *Estimating effect sizes.* Which effect size metric should be used? What are the best ways to convert study statistics into effect sizes? How can extraneous influences on effect size magnitude best be controlled?
6. *Analyzing the database.* How should the effect size data be analyzed statistically? Which of the available meta-analytic frameworks for statistical analysis is most appropriate? What sorts of statistical models are appropriate? How can the tests associated with these models be interpreted? How can statistical outliers among the effect sizes be located and treated?
7. *Presenting, interpreting, and disseminating the results.* What information about the studies should be presented? Which meta-analytic models should appear? What are the best techniques for displaying the meta-analytic results? What knowledge accrues from the synthesis?

How do the meta-analytic results reflect on the theoretical analysis? Has the synthesis uncovered important areas that warrant future research? Has it revealed novel hypotheses that should be tested in new primary research?

Conceptual Analysis of the Literature

The initial conceptual exploration of a research literature is critical because these ideas affect the methods that follow, such as the criteria for including and excluding studies. The first conceptual step is to specify, with great clarity, the phenomenon under review by defining the variables whose relation is the focus of the review. Ordinarily, a synthesis evaluates evidence relevant to a single hypothesis that is defined as a relation between two variables, often stated as the influence of an independent variable on a dependent variable (e.g., the effects of ego depletion on self-control, synthesized by Hagger, Wood, Stiff, and Chatzisarantis, 2010). Moreover, a synthesis must take study quality into account at an early point to determine the kinds of operations that constitute acceptable operationalizations of these conceptual variables. Typically, studies testing a particular hypothesis differ in the operations used to establish the independent and the dependent variables. If the differences in studies' operations can be appropriately judged or categorized, analysts can probably explain some of this variability using these differences as moderator variables.

The research problem's history and its typical studies are essential to this conceptual analysis. Theoretical articles, earlier reviews, and empirical articles should be examined for their interpretations of the phenomenon under investigation. Authors' theories or even their more informal insights may suggest moderators of the effect that could potentially be coded in the studies and examined for their explanatory power. If scholars have debated different theories, the synthesis should be designed to address them, if possible.

The most common way to test competing explanations is to examine how the findings pattern across studies. Specifically, a theory might imply that a third variable should influence the relation between the independent and dependent variables: The relation should be larger or smaller with a higher level of this third variable. Treating this third variable as a potential moderator, the analyst would code the studies for their status on the moderator. This meta-analytic strategy, known as the *moderator variable* or *effect modifier approach*, is analogous to the examination of interactions with primary-level data (see the section on Estimating Effect Sizes). However, instead of testing the interaction

within one study's data, the metaanalysis tests whether the moderator affects the examined relation across the studies included in the sample. Such an analysis determines *when* the magnitude or sign of the relationship varies. Using this strategy, Malle (2006) found that the tendency to explain one's own behavior with situational causes and others' behavior with personal causes holds only for negative events; the opposite asymmetry holds for positive events.

In addition to this moderator variable approach, other strategies have proven to be useful. In particular, a theory might suggest that a third variable serves as a mediator of the critical relation because it conveys the causal impact of the independent variable on the dependent variable (see Judd, Yzerbyt, & Muller, Chapter 25 in this volume; Shadish, 1996). If at least some of the primary studies have evaluated this mediating process, mediator relations can be tested within a meta-analytic framework by performing correlational analyses that are an extension of path analysis with primary-level data. Using such techniques, Albarracín, Johnson, Fishbein, and Muellerleile's (2001) examination of 96 independent studies showed that, consistent with reasoned action approaches, intentions generally mediated the influence of attitudes, subjective norms, and perceived behavioral control on action.

Setting Boundaries for the Sample of Studies

In beginning a metaanalysis, the reviewer should consider whether all possible tests of a relationship should be included. This decision is important because the inferential power of any metaanalysis is limited by the methods of the studies that it integrates. To the extent that all (or most) of the reviewed studies share a particular methodological limitation, any synthesis of these studies would be limited in this respect. For example, a synthesis of correlational studies will produce only correlational evidence about the association in question. Yet if the critical hypothesis were tested with true experiments, defined by one or more manipulated independent variables and the random assignment of participants to conditions, the metaanalysis would gauge the causal effect of the independent variables on the dependent variable across the studies reviewed. Nevertheless, in all metaanalyses, most relations between moderator variables and the effect of interest are correlational and therefore causally ambiguous. For example, Koenig, Eagly, Mitchell, and Ristikari (2011) found that, across three research paradigms, the cultural masculinity of the leader stereotype has decreased over time. Effects of year of publication, like many other study characteristics, can be difficult to interpret because of potential confounds with other variables (e.g.,

cultural change or change in methods).

Moderator tests can yield stronger causal claims if the moderator reflects within-studies manipulations. In such cases, random assignment of participants to levels of the moderator in the primary studies makes it less likely that confounds were associated with the moderator. In this strategy, the results of each study are divided to produce separate effect sizes within levels of the moderator. For example, Baas, De Dreu, and Nijstad (2008) showed that creativity was enhanced more by positive moods than by neutral ones; moreover, mood valence was experimentally manipulated in most of the studies. If an analysis were limited to the studies that contained this manipulation, any moderation could be more confidently attributed to the manipulated variable, barring confounds with other variables.

In deciding whether some studies may be insufficiently rigorous to include in the metaanalysis, a reviewer should take into account methodological standards within the research area. Although a large number of potential threats to methodological rigor have been identified (see Brewer & Crano, Chapter 2 in this volume; Campbell & Stanley, 1963; Shadish, Cook, & Campbell, 2002; Valentine, 2009), there are few absolute standards of study quality. For example, there are hundreds of scales purporting to gauge methodological quality (Deeks, Dinnes, D'Amico, Sowden, Sakarovich, Song, Petticrew, & Altman, 2003). Moreover, in practice, the characteristics considered essential to ensure high study quality vary widely across research areas. In some literatures, it is known that a certain method (e.g., a measure or a manipulation) yields seriously flawed results; if so, an analyst might eliminate studies that used this method. Indeed, one possible strategy is to omit obviously flawed studies to restrict the synthesis to studies of high quality, a practice known as *best-evidence synthesis* (Greenwald & Russell, 1991).

Another option is to attempt to correct the effect sizes for certain methodological biases (see the section on Correcting Effect Sizes for Bias). Retaining potentially flawed studies and representing their quality-relevant features in the coding scheme is another defensible strategy, given that methods always contain some degree of error. For example, if a given variable was not manipulated or assessed uniformly across the studies, a coding of the variable's quality (e.g., its reliability) may predict effect size magnitude. More generally, metaanalyses should examine whether variant methods yield differing findings (for an example, see Heinsman & Shadish, 1996; see also Moyer & Finney, 2002).

In addition to study quality, many other considerations enter into setting the boundaries of a research literature. Boundary-setting forces reviewers to weigh conceptual and practical issues, which are particularly acute in literatures featuring a variety of methods. Sometimes boundaries include only studies that are relatively homogeneous methodologically (e.g., only experimental studies), and sometimes boundaries encompass different methods (e.g., both experimental and correlational studies). In general, boundaries should be wide enough to allow the testing of interesting hypotheses about moderator variables. Yet if very diverse methods are included, some moderator variables may exist only within particular methods (e.g., participants' organizational status exists only within studies conducted in organizations). In general, including a wide variety of methods might make a metaanalysis unwieldy. In such instances, meta-analysts may divide a literature into two or more reviews, each addressing a different aspect of a broad research question.

If the boundaries of a metaanalysis are too wide, researchers may be the targets of what is known as the “apples and oranges” critique (Glass et al., 1981) – that is, combining studies that used markedly different methods. Methodologists have been generally unsympathetic to this criticism because they regard it as the task of the meta-analyst to examine whether differences in methods produce consequential differences in study outcomes. This demonstration is achieved by dividing studies into various categories or ranges, as we discuss in the section on Analyzing the Meta-Analytic Database. Of course, metaanalyses that fail to consider moderators can warrant the criticism of ignoring the possible effects that diverse methods have on study outcomes.

Analysts often set the boundaries of the synthesis so that the methods of included studies differ dramatically only on critical moderator dimensions. If other extraneous dimensions are thereby held relatively constant across the reviewed studies, moderator variable analyses can be more clearly interpreted. Meta-analysts proceed by dividing studies based on the moderator variable, where possible, and analyzing the effect of interest within the levels of the moderator (or treating such moderators as continuous variables). Such designs appear frequently in social and personality psychology. For example, because argument quality moderates the effects of involvement on message-based persuasion, Johnson and Eagly (1989) calculated involvement effect sizes within the levels of quality.

Meta-analysts should include all studies or portions of studies that satisfy the selection criteria. If some studies meeting preliminary criteria established

conditions that are judged to be extremely atypical (e.g., mentally disabled or ill participant populations), the selection criteria may be modified to exclude them. Developing selection criteria often continues as meta-analysts examine more studies and thereby discover the full range of research designs that have investigated a particular hypothesis.

One issue that generally arises when setting boundaries is whether to include unpublished studies (Rothstein & Hopewell, 2009). Although these studies are usually more difficult to access, their omission typically biases the review's findings in favor of larger effects (e.g., Dickersin, 1997; Johnson, Scott-Sheldon, & Carey, 2010; Lipsey & Wilson, 1993). The frequent omission of nonsignificant findings from the research record is most likely responsible for the so-called *decline effect* (Schooler, 2011), whereby the strength of findings supporting a particular hypothesis decreases after initially appearing robust. Moreover, the withholding of nonsignificant findings from publication appears to be a widespread practice that can compromise the validity of many published effects (Francis, 2012; Ioannidis, 2005). In a discussion of unpublished studies, Rosenthal (1979) referred to them as producing “a file-drawer problem” because they may be buried in researchers’ file drawers and therefore inaccessible to reviewers. In fact, surveys of researchers suggest that as much as two-thirds of the studies that are conducted are never published (Cooper, DeNeve, & Charlton, 1997; Rotton, Foos, Van Meek, & Levitt, 1995). Of course, many additional factors affect studies’ publication status (e.g., author productivity; Sommer, 1987). A partial solution to the problem of published literatures that are biased in favor of hypotheses is to seek studies that are reported in dissertations and master's theses and as poster sessions and talks at conferences. These studies are less likely to be screened for statistical significance than studies published in journals. Meta-analysts can also ask researchers in an area if they have additional, unpublished data sets that they can share.

Given these considerations, every effort should be made to obtain unpublished studies. The goal of metaanalysis is to describe the *universe* of studies on a topic, or at least an unbiased sample of that universe (White, 2009). Disregarding this goal compromises the validity of the metaanalysis as a representation of the research literature. Ironically, a meta-analyst would not even learn that this unpublished literature exists without searching for it. Another benefit of including unpublished studies is that they enlarge the number of studies in the metaanalysis, thereby increasing statistical power to estimate mean effect sizes and to detect moderators of effect sizes.

Regardless of studies' publication status, analysts should judge them against a set of inclusion and exclusion criteria and code their quality-relevant features. Uniform implementation of these procedures helps circumvent the potential criticism that unpublished studies are generally of unacceptable quality because of the absence of peer review. Rather than merely assume (perhaps incorrectly) that unpublished studies are of inadequate quality, a meta-analyst should remove all studies, published or unpublished, that do not meet the review's quality criteria and code the remaining studies on quality-relevant study characteristics (e.g., reliability of measures).

A further decision that often arises is whether the sample of studies should be restricted to one country or culture. The reasoning that encourages sampling unpublished studies also encourages sampling studies from all countries and cultures. Moreover, including such studies increases the inclusiveness of the metaanalysis by permitting an analyst to answer questions about the generality of the studied effect across diverse cultures. Indeed, it seems meritorious for metaanalyses with large enough samples of studies to conduct such tests routinely. For example, Bond and Smith (1996) found that conformity in Asch-style line-judgment experiments was more marked in collectivistic than in individualistic cultures (although the conformity effect was significant within both types of cultures). Yet, in many research literatures, it may not be possible to address this issue meta-analytically because only a very small number of studies are available from countries other than the one in which the research paradigm first appeared (e.g., Eagly, Makhijani, & Klonsky, 1992; Lee et al., 2011). Therefore, as a general rule, studies from multiple cultures should be included in the sample if they are available in at least modest numbers. Although computer applications (e.g., Google Translate) can help overcome foreign language barriers, knowledge of a culture's practices can be crucial to coding such studies accurately. Therefore, meta-analysts should typically seek the assistance of native or other highly skilled speakers of the foreign languages represented in the included studies (e.g., Pettigrew & Tropp, 2006).

A final issue is the completeness with which very large research literatures are reviewed. Some literatures are so enormous that including *all* studies would be impractical. In these instances, meta-analysts might take a random sample of the entire research literature (Card, 2012), with sample size guided by statistical power considerations (Cafri, Kromrey, & Brannick, 2009; Valentine, Pigott, & Rothstein, 2010).² Specifically, a meta-analyst would list all the studies in the pertinent literature, decide how many would make a sufficient sample, and

randomly select this number of studies. An example of such sampling is Rosenthal and Rubin's (1978) metaanalysis of the interpersonal expectancy effect literature.

Locating Relevant Studies

Because including a large number of studies generally increases the value of a quantitative synthesis, it is important to locate as many studies as possible that might be suitable for inclusion. When a literature consists of findings whose presence in reports cannot necessarily be discerned from reading titles and abstracts, a reviewer may have to retrieve all studies in the general research area to identify the finding of interest. For example, Kotov, Gamez, Schmidt, and Watson (2010) screened 7,156 abstracts of studies on traits and anxiety, depression, and substance use; 175 studies fit their inclusion criteria.

Reviewers are well advised to err in the direction of being overly inclusive in their searching procedures. As described elsewhere (e.g., Cooper, 2010; Johnson & Boynton, 2008; Lipsey & Wilson, 2001; White, 2009), there are many ways to find relevant studies; ordinarily, analysts should use all of these techniques. Unfortunately, computer searches of databases such as PsycINFO and Google Scholar seldom locate all of the available studies, although such searches are extremely useful. There are many other databases aside from the most familiar aforementioned ones. Some of these databases cover literature primarily in English (e.g., ProQuest Dissertations and Theses, Web of Science, Sociological Abstracts, MEDLINE, ABI/Inform Global, ERIC). Other databases contain primarily studies published in foreign languages (Psicodoc for Spanish and Portuguese; PSYNDEX for German). Also, other nations maintain databases of dissertations (e.g., Index to Theses and Electronic Theses Online Service, United Kingdom and Ireland; Deutsche Nationalbibliothek and Dissonline, Germany; DART-Europe, pan-European portal for dissertations and theses; China Doctor Dissertations Database). Finally, conference papers and other types of unpublished papers appear in PsycEXTRA (from the American Psychological Association) and ERIC. These databases thus provide partial access to the fugitive literature of unpublished studies (Rothstein & Hopewell, 2009). Databases also increasingly afford full-text searches, which can be very important for literatures in which the focal comparison is less likely to appear in abstracts (e.g., comparison of cooperative behavior of women and men as reviewed by Balliet, Li, Macfarlan, & Van Vugt, 2011). Librarians can provide helpful advice to novice searchers, and many databases offer excellent tutorials

(Reed & Baxter, 2009).

Finally, to enable evaluation of search procedures as well as their replication, the review should describe in detail its methods of locating studies, including the names of the databases that were searched, and for each database the time period covered and the keywords used. Reviewers should also describe their inclusion and exclusion criteria and provide a rationale for these criteria, consistent with metaanalysis reporting standards (MARS; American Psychological Association, 2008). More comprehensive standards (e.g., PRISMA; Moher, Liberati, Tetzlaff, & Altman, 2009) include other features, such as a chart describing the flow of study reports into the metaanalysis and a listing of excluded as well as included studies.

Study Characteristics.

In conceptualizing the metaanalysis, reviewers have usually developed ideas about the study characteristics that should be coded. The most important of these characteristics are potential moderator variables that may account for variation among the studies' effect sizes. It is also important to consider whether studies that differ along a critical moderator dimension also differ on other dimensions. Because such confounds could produce interpretational difficulties, coding these additional characteristics potentially permits a metaanalysis to determine which variables explain unique variation in predicting effect size magnitude and which do not. Finally, it is also important to code the studies for numerous other characteristics such as their date of publication and participant population, even if these characteristics are not expected to account for variation in studies' outcomes (Lipsey, 2009), because such features help set an interpretative context for the review.

Study characteristics may be either continuous or categorical. Variables on a *categorical* metric consist of a discrete number of values that reflect qualitative differences between those values. For example, among the categorical study characteristics that Freund and Kasten (2012) coded in a metaanalysis of the validity of self-estimates of cognitive ability were ability type, order of self-estimate and ability test, and gender of participants. Variables on a *continuous* metric consist of values that exist along ratio, interval, or ordinal scales (see Wilson, 2009 for examples).

Some important features of studies are difficult to code accurately by reading study reports. For example, in a metaanalysis on sex-related differences in aggression, Eagly and Steffen (1986) wished to determine whether women and

men differed in how unfavorably they perceived aggressive acts. Therefore, they asked female and male students to rate the extent to which each such act would produce harm to the target of aggression, guilt and anxiety in oneself as the aggressor, and danger to oneself. From these ratings Eagly and Steffen estimated sex differences in these students' perceptions of the aggressive acts and related these scores to the effect sizes that represented sex differences in aggressive behavior. In other instances, experts' ratings could be obtained based on their reading of the method sections of the reports or of the actual stimulus materials used in the studies (e.g., Johnson & Eagly, 1989; Marcus-Newhall, Pedersen, Carlson, & Miller, 2000). Similarly, in a review of the involvement and persuasion literature, Johnson and Eagly (1989) provided undergraduate judges samples of the arguments these studies had used and asked them to rate them in terms of their strength in supporting the message position. Such operations help assess dimensions that can prove important in moderator analyses.

Convergent evidence of the reliability and validity of the judges' ratings used by these methods is desirable, because these judges function only as observers of studies' methods. Interjudge reliability estimates can be calculated (e.g., Marcus-Newhall et al., 2000). In addition, the validity of judges' ratings of manipulation effectiveness can be estimated by comparing them with effect sizes representing the manipulation checks present in the studies (e.g., Bettencourt & Miller, 1996; Miller, Lee, & Carlson, 1991).

Reliability of Coding.

Given the importance to metaanalyses of accurate coding of the included studies, two or more individuals should perform the coding independently, followed by the calculation of an appropriate index of interrater reliability (such as the intraclass correlation or Cohen's, 1960, *kappa*; Orwin & Vevea, 2009). In most cases, disagreements can be resolved by discussion, or perhaps by averaging. Given that coding can be extremely time consuming, an alternative is to conduct dual coding on only a subset of studies, and if reliability is high, do only single coding on the remaining studies (Card, 2012). However, random sampling should determine which studies enter the initial sample of studies to be double-coded. Then, once reliability is established, studies should be chosen at random for double-coding (and included in the final reliability calculations). The better procedure, if feasible, is to double-code all studies.

Cultural and Social Structural Characteristics.

Although meta-analysts rely mainly on information in the source reports, they often incorporate information available elsewhere. Such information ranges from physical dimensions of social milieus to descriptions of social collectives such as organizations, communities, and nations. For example, Mullen and Felleman (1989) learned what specific dormitories had been studied in studies of crowding and then obtained from college administrators blueprints that allowed them to gauge physical features that were relevant to crowding effects. Similarly, Eagly, Johannesen-Schmidt, and van Engen's (2001) synthesis of sex differences and similarities of leadership styles obtained data from the U.S. Bureau of Labor Statistics and other sources to estimate the distribution of the sexes in studies' leadership roles when that information was missing from the reports.

Many additional databases relevant to social and personality phenomena track trends over decades or even centuries. For example, Gapminder (2012) tracks nation-level indicators on hundreds of dimensions (e.g., economic and health statistics). Among the databases that make U. S. survey data available are the American National Election Studies (2012) and the General Social Survey (2012). Many other nations and collectives (e.g., International Social Survey Programme, 2012) conduct similar opinion surveys. Hofstede's (2001) and others' surveys on cultural dimensions such as individualism, uncertainty avoidance, and masculinity are available for many nations (Taras, Kirkman, & Steel, 2010). The Cingranelli-Richards Human Rights Project (2012) gauges government respect for human rights across most nations. The World Values Survey (2012) compiles political and sociocultural indicators for many nations. The United Nations Statistics Division (2012) offers economic and sociopolitical data, as do the International Labor Organization (2012) and the World Bank (2013).

Estimating Effect Sizes in Individual Studies

To be included in the metaanalysis, a study must report a quantitative test of the hypothesis under scrutiny. In theory, each study j provides an observed estimate, T_j , of the underlying population phenomenon, θ . Hence, an observed study result is not the “truth” but an estimate of it. In general, past metaanalyses in personality and social psychology have emphasized two-variable quantitative tests, such as how maternal employment relates to children's achievement (Goldberg, Prause, Lucas-Thompson, & Himself, 2008). Other metaanalyses have used the arithmetic means of one or more variables as effect sizes – for example, how much well-being, burnout, and anxiety are present in particular

nations (Fischer & Boer, 2011). This section considers two-variable effect size indexes, otherwise known as indexes of association, and the following section addresses arithmetic means.

Effect Size Indexes of Association.

There are many effect size indexes that gauge associations between two variables, as Table 26.1 shows. The table indicates that the measurement features of the variables in question guide the choice of effect size and the particular effect size index. As a general principle, if two or more studies report any one of Table 26.1's effect size metrics, they can be meta-analyzed, although all results must be converted to a single metric.³ In addition to an effect size index T_j for each study j , the sampling error associated with each study's effects must be estimated or recorded because it is used in all analyses. In social and personality psychology, because a diversity of measures appears to be the rule, analysts have nearly always used standardized effect size indexes, especially the standardized mean difference and the correlation coefficient. These effect sizes yield a common metric for comparing studies' findings.

Table 26.1. Potential Two-Variable Effect Sizes Dependent on the Measurement Features of the Two Variables (adapted from Johnson & Boynton, 2008).

Nature of Second Variable	Nature of First Variable		
	Continuous	Ordinal	Categorical
<i>Continuous</i>	<ul style="list-style-type: none"> • Pearson correlation (r) • Standardized regression slopes (β) • Unstandardized regression slopes 	<ul style="list-style-type: none"> • Biserial correlation (r_b) 	<ul style="list-style-type: none"> • Standardized mean difference • Unstandardized mean difference • Point-biserial correlation (r_{pb})
<i>Ordinal</i>		<ul style="list-style-type: none"> • Spearman correlation (ρ or ρ) • Tetrachoric correlation (r_{tet}) 	<ul style="list-style-type: none"> • Rank-biserial correlation
<i>Categorical</i>			<ul style="list-style-type: none"> • Phi coefficient (ϕ) • Odds ratio (OR) • Risk ratio (RR) • Risk difference (RD)

Note: (a) Whether a variable is “first” or “second” is arbitrary. (b) “Categorical” assumes two discrete categories (e.g., male vs. female or experimental vs. control group), but it is of course possible to have more than two categories. (c) Any continuous or ordinal variable(s) could artificially be placed in a

coarser category. (d) Some forms of effect size have sub-types not listed here (e.g., standardized mean difference can gauge either the means of two independent groups or of two time points for a single group).

Table 26.2 provides equations for the most commonly used forms of the standardized mean difference, the product-moment correlation coefficient, and the logged odds ratio. The table also highlights the systematic biases of estimates of effect sizes that are typically corrected in analyses. In addition, this table notes changes in the naming conventions for standardized mean differences. For example, Hedges's d (line 2) also has been labeled g^* and Hedges's g . Hedges (1981) developed this particular index of T specifically to apply to between-groups comparisons at a single point in time, providing proofs and documentation pertaining to this type of comparison. (Other sources consider complexities such as adjusting baseline differences between groups or gauging their change over time; e.g., Becker, 1988; Table 26.2, line 3). Consequently, the term “Hedges's d ” should be restricted to the comparison specified in Table 26.2's line 2. The same principle holds regarding the other indexes of T .

Table 26.2. Common Two-Variable Effect Size Equations, Inverse Variance, and Usage Notes

Number	Effect Size	Equation(s)	Inverse Variance	Terms	Classic Citation and Notes									
Standardized Mean Difference														
<i>Two-group comparison</i>														
1	Cohen's d	$d = \frac{M_A - M_B}{SD_P}$	Not formally defined	M_A = mean for group A; M_B = mean for group B; SD_P = pooled standard deviation	Cohen's (1969) d is often called "uncorrected effect size" or g to distinguish it from Hedge's (sample-size corrected) d .									
2	Hedges's d	$d = J(m) \times$ Cohen's d , where $J(m) \approx 1 - \frac{3}{4m - 1}$	$\frac{2(n_a + n_b)n_a \times n_b}{2(n_a + n_b)^2 + n_a n_b d^2}$	$m = n_A + n_B - 2$ n_a = sample size for group a; n_b = sample size for group b	Hedges's (1981) d is often termed "Hedges's g " and sometimes g^* , where the asterisk implies the sample-size correction, $J(m)$. In the inverse variance equation, d is Hedges's d .									
<i>One-group temporal comparison</i>														
3	Becker's d	$d = \frac{M_{Pre} - M_{Post}}{SD_{Pre}}$	$\frac{2N}{4(1 - r) + d^2}$	M_{Pre} = pretest mean; M_{Post} = posttest mean; SD_{Pre} = pretest standard deviation; r = correlation between pretest and posttest; N = sample size	Becker (1988)									
Correlation between two variables														
4	Pearson's product-moment r	$r = \frac{\sum_{i=1}^N z_{X_i} z_{Y_i}}{N}$	Not formally defined	z_{X_i} and z_{Y_i} = standardized forms of X and Y being related for each case i	Pearson (1895)									
5	Correction to Pearson's r	$\tilde{G}_{(r)} \cong r + \frac{r(1 - r^2)}{2(N - 3)}$	Not formally defined		Rarely used because bias is small when $n > 20$.									
6	Fisher's r -to- z transform	$z_r = \frac{1}{2} \log_e \frac{1 + r}{1 - r}$	$N - 3$	\log_e = natural logarithm	Fisher (1921)									
7	Fisher's z -to- r transform	$r = \frac{e^{(2Z_r)} - 1}{e^{(2Z_r)} + 1}$	Not formally defined	e = base of the natural logarithm	Fisher (1921)									
Odds ratio														
8	Logged odds ratio (OR)	$LOR = \log_e \left(\frac{ab}{bc} \right)$	$\frac{abcd}{ab(c + d) + cd(a + b)}$	\log_e = natural logarithm Observed cases in a 2×2 contingency table: <table border="1"><tr><td></td><td>+</td><td>-</td></tr><tr><td>+</td><td>a</td><td>b</td></tr><tr><td>-</td><td>c</td><td>d</td></tr></table>		+	-	+	a	b	-	c	d	A. W. F. Edwards (1963), J. H. Edwards (1957)
	+	-												
+	a	b												
-	c	d												
9	Transform of logged odds ratio to OR	$OR = e^{LOR}$	Not formally defined	e = inverse natural logarithm function	Is used to convert the LOR back into its original units for purposes of display and interpretation.									

Note: The inverse variance is provided only for fixed-effects assumptions. For random-effects assumptions, see the text.

The Direction of Effect Sizes Gauging Associations.

No matter the type of T used in a metaanalysis, its direction must be maintained consistently across the included studies by making T positive or negative so that studies with opposite outcomes have opposing signs. Ordinarily, a positive sign is given to outcomes in the expected, hypothesized, or typical, direction for the metaanalysis as a whole, whereas the negative sign is given to outcomes that reverse this direction. Only a relation that is exactly null would have no sign, because a standardized mean difference effect size (or r) would be 0.00 (and the Odds Ratio would be 1.00).⁴ Illustrating this practice is Kite and Whitley's (1996) metaanalysis of sex-related differences in attitudes toward homosexuals, in which the expected direction of the findings was that women would evaluate homosexuals more positively than do men. Therefore, the positive sign for effect sizes indicated that women's evaluations were more positive than men's, and the negative sign that men's evaluations were more positive than women's. Alternatively, when experimental groups are compared with control groups, differences in favor of the experimental group might be given a positive sign, and differences in favor of the control group given a negative sign. Finally, metaanalyses may examine omnibus T s, such as multiple R , which gauges the amount of variance explained in a dependent variable attributable to more than one predictor variable; such T s take only positive signs.

Multiple Reports from Individual Studies.

When a given study provides multiple reports of the relation of interest, the analyst must decide whether to average the effect sizes to represent the study with a single effect size estimate or to treat them as separate estimates. To preserve the independence of the effect sizes in a metaanalysis, each must come from a different study. That is, the participants whose data contribute to a given effect size must not contribute to any other effect sizes in the analysis.⁵ Therefore, the analyst would ordinarily average multiple effect sizes calculated from a single study. Instead of or in addition to averaging, an analyst may wish to investigate whether the results of the studies varied depending on the different operations by which their dependent variables were defined. For this purpose, the preservation of the separate effect size estimates made within individual studies may enable subsequent analyses examining whether the operations produced differences in the effect sizes. For example, in a metaanalysis of sex differences in leaders' effectiveness, Eagly, Karau, and Makhijani (1995) analyzed effect sizes according to the identity of the raters who provided the

effectiveness measure and the basic type of measure (e.g., objective vs. subjective). Although many individual studies contributed several effect sizes to these analyses, each study's effect sizes were subsequently aggregated into a single study-level effect size that was used in additional analyses that did satisfy the assumption that effect sizes are independent. Analyses using multiple effect sizes from single studies can be informative even though they violate the assumption of independence of the effect sizes and thus can make statistical tests more liberal than they ought to be.

When a study examined the focal relation within levels of another variable, effect sizes may be calculated within these levels as well as for the study as a whole. How seriously the use of such within-level effect sizes violates the independence assumption depends on whether these levels were created on a within-subjects or a between-subjects basis. If the same participants took part at all levels of the variable (i.e., a within-subjects variable), the effect sizes would be highly dependent. The effect sizes would also be dependent if one control group served as a comparison for more than one treatment group. Even if the participants at the different levels were not the same individuals, the effect sizes would be dependent because they came from the same study, which was carried out under conditions existing in a particular place at a particular point in time (Hedges, 1990). For example, effect sizes might be calculated separately for the male and female participants of studies to enable examination of sex-related differences in the relation (e.g., Koenig et al., 2011), even though these effect sizes would not be independent.

Precision of Reported Statistical Information.

Reports may contain more than one form of statistical information that could be used to calculate a given effect size. Some of these should converge within rounding error. For example, F -tests or t -tests should produce the same T as do the means and standard deviations that underlie them. The analyst should compute the effect size from both such sources to make sure that the results agree. As long as the effect sizes are similar, they should be averaged. If the effect size estimates are dissimilar, there may be errors in the information reported or the analyst's calculations. Sometimes inspection of the report's quantitative information for its internal consistency suggests that one form of the information is more accurate.

Similarly, for many reasons, some source reports contain less than desirable amounts of information for estimating T s, especially when T is gauged as a

standardized mean difference. Some routes to estimating effect sizes merely require a great deal of effort on the part of the analyst (e.g., reanalyzing raw data found in an appendix of a dissertation). In other instances, deriving an effect size may require the application of several nonroutine techniques in sequence. (We provide some of these strategies in the Appendix.) Each metaanalysis poses statistical challenges that may call for novel solutions.

Meta-analysts should contact studies' authors, if possible, to acquire essential information that is not included in a report. In our experience, cordial invitations to authors have produced moderate success rates (e.g., 40%). Obtaining such information allows the report to be adequately represented; failing to obtain the needed information renders the metaanalysis less comprehensive and potentially less representative. Finally, a lack of statistical detail in reports does not necessarily reflect their authors' oversights, errors, or poor methods. Rather, omissions generally occur because the authors' goals differed from those pursued in a subsequent metaanalysis. For example, a small sex-of-employee effect on job performance might have warranted only a brief acknowledgement of its nonsignificance, but for a metaanalysis on this subject (e.g., Roth, Purvis, & Bobko, 2012), such findings are crucial.

Dealing with Nonreported Results.

Reports that describe the effect of interest merely as “nonsignificant” are highly problematic in metaanalysis (Bushman & Wang, 1996). It is common to represent such effects as though they are exactly null (e.g., $d = 0.00$), but such estimates are obviously crude. If the N in the study was small, its actual effect size could be quite large, yet not significant. Introducing such effect sizes into a metaanalysis as though they were null biases a mean effect size toward the null (Schmidt, 1996); when these studies actually have results in the opposite direction, then assuming a null value is also unsatisfactory. Especially if many such reports exist in a literature, it may be advisable to conduct analyses with and without these 0.00 values.

At the synthesis stage of a metaanalysis, one way to incorporate imprecisely reported results, including those described as nonsignificant, is to use so-called “vote-counting procedures” to summarize findings (Bushman & Wang, 1996; Darlington & Hayes, 2000). In these procedures, rather than using effect size estimates to represent the studies' outcomes, an analyst examines how many studies obtained a result in the hypothesized direction or how many obtained a significant result in this direction. Because the strategy relies only on findings'

directions or significance levels, it allows an analyst to include even the imprecisely reported nonsignificant results. More formally, calculating what is sometimes called the “sign test” determines the exact p of the observed distribution of positive and negative outcomes (or one more extreme), given that the probability of obtaining a positive result is .5, according to the null hypothesis, which specifies that half of the results should be positive and half negative following the binomial distribution. This probability can be calculated by standard statistics packages or spreadsheet software. An analyst can also use the binomial distribution to calculate a p -value for obtaining the observed distribution of significant positive findings versus other findings (nonsignificant and reversed), given that the probability of obtaining a significant result in one tail of the distribution is .025, according to the null hypothesis and assuming .05 for two-tailed significance testing. The p -values associated with the proportion of the studies that have a positive direction or that produced a significant positive result can be used to estimate a mean effect size for a sample of studies. These estimated effect sizes can then be compared to the exact mean effect size based on the studies that permitted this calculation (Bushman & Wang, 1996). For example, Wood (1987) used these techniques to estimate the mean effect size for sex-related differences in group performance because many of the studies did not permit an effect size to be estimated. Of course, it is much better to calculate the mean effect size by averaging effect sizes from individual studies when the majority of studies permit this strategy.

Reliability of Effect Size Calculations.

At least two analysts should compute effect sizes independently for each of the studies and then compare solutions and resolve discrepancies. Given the complexity of many research designs and the ambiguity of some research reports, errors of effect size estimation are not uncommon. Moreover, sometimes one analyst may discover an indirect route to computing an effect size that is missed by a second analyst. Calculations by two or more analysts minimize such errors and omissions (see the section on Reliability of Coding).

Correcting Effect Sizes for Biased Methods.

In addition to correcting the raw g and r for their inherent bias as estimators of the population effect size (see prior subsection on Effect Size Indexes), analysts may correct for many other biases that accrue from the methods used in each study. For example, as the reliability of a measure increases (and its

measurement error therefore decreases), its relations with other variables will also increase (Cronbach, 1990). Increased measurement error decreases a measure's ability to predict another variable. Corrections for measurement unreliability and other forms of error or bias allow estimation of the strength of a relation absent such artifacts. In their presentations of such corrections for independent and dependent variables, Hunter and Schmidt (2004) and their colleagues (e.g., Schmidt, Le, & Oh, 2009) explained how to implement corrections for measurement error, artificial dichotomization of a continuous variable, imperfect construct validity, and range restriction. In theory, correcting for such errors permits a more accurate estimation of the true population effect size.

These corrections are quite popular in industrial and organizational psychology (e.g., Chiaburu, Oh, Berry, Li, & Gardner, 2011). They have seldom been used in social psychological metaanalyses because in most research areas relatively few studies include the information that would be required to perform the corrections (e.g., reliability or validity statistics). Nevertheless, meta-analysts may perform such corrections in research literatures in which reliabilities and other relevant information are routinely provided.

When meta-analysts do implement these corrections, the resultant corrected mean effect size yields an idealized estimate of the magnitude of the population effect rather than an estimate of the relation that is reported in a typical study if the corrections were not implemented. Nonetheless, because the correction procedures assume that the different biases are uncorrelated, the bias-adjusted corrections can yield irrational effect sizes (e.g., correlations larger than 1.00; Rosenthal, 1991). Therefore, analysts should consider their goals when deciding whether to use such corrections. If the goal is to estimate the effect size that would exist if there were no contamination by artifacts of measurement, the corrections would be desirable. In contrast, if the goal is to show how large a relation is in practice, then the corrections would be less useful.⁶

Regardless of whether these corrections are implemented, various biases may enter into studies' effect sizes. Consider that effect size estimates are a ratio of signal to noise, like all inferential statistics. For example, in a between-groups design, the signal is the difference in means, and the noise is the pooled standard deviation. Methodological factors can influence the effect size through their impact on signal, noise, or both factors. If two identical studies are conducted and one controls for noise that the other study does not (e.g., by statistically controlling for an individual difference characteristic), the first study will have a

smaller error term (standard deviation), and the effect size will be larger for the first than the second study. To minimize this type of variation in effect sizes, meta-analysts should equate as much as possible the comparisons that the studies yield, so that the effect sizes are not influenced by differing statistical operations. For example, one such recommendation is that in metaanalyses of experimentally manipulated effects, analysts return irrelevant individual difference factors to the error term if they were included in the analysis in only some of the included studies. Reconstituting the error term in this way would not be necessary if the variable in question were controlled in all of the studies in the review. Similarly, many contemporary statistics already invoke corrections. For example, causal models with a latent variable structure effectively correct for unreliability and invalidity. Consequently, including results from such studies along with studies without latent variable structures introduces methodological noise across a literature. One method to reduce this influence is introducing the Hunter-and-Schmidt bias corrections to studies that lack the corrections (Card, 2012).

Additional problems can arise from the inclusion of studies that used within-subjects designs. For example, a researcher might have implemented a within-subjects design that required each participant to judge two objects along the same dimension. Such multiple assessments can produce many complications, including carryover, priming, and contrast effects (Smith, Chapter 3 in this volume). In analyzing such data, researchers nearly always use a repeated-measures inferential statistic that removes within-subjects variation from the error term. Consequently, these tests are more statistically powerful than those produced by a comparable between-subjects design (Dunlap, Cortina, Vaslow, & Burke, 1996; Morris & DeShon, 2002). If the meta-analyst uses these within-subjects error terms to calculate effect sizes, it is likely that these effect sizes will be larger than those based on standard deviations pooled from the cells of the design (e.g., Kite & Johnson, 1988; for an exception, see Symons & Johnson, 1997). Some sources recommend not mixing effect sizes from these two types of designs in the same analysis (e.g., Lipsey & Wilson, 2001), but others suggest using type of design as a moderator variable (e.g., Card, 2012). A growing convention is to estimate within-subjects cases using a between-subjects approximation (Becker, 1988).

Although it is unrealistic for analysts to take into account all potential sources of bias in a metaanalysis, they should remain aware of potential biases within their research literature. Some of these biases can be corrected in the process of computing the effect sizes. Others can be examined empirically for their

influence on studies' results. Still others can be eliminated by narrowing the boundaries of the literature under investigation to exclude biased studies. When it is not possible to control a bias in some fashion, analysts should consider what influence it might have on their findings and interpret the results accordingly.

Using Arithmetic Means to Gauge a Quantity's Magnitude

In the last 15 years, some meta-analysts in personality and social psychology have conducted metaanalyses by analyzing arithmetic means from studies as their estimate of T . With such strategies, analysts examine how low or high a sample scored on a certain criterion and model these outcomes using information about the samples (e.g., gender, recruitment strategies) and their milieus (e.g., economic success of women). For example, Twenge and her colleagues have examined temporal trends in U.S. samples in terms of levels of such variables as anxiety (Twenge, 2000), depression (Twenge & Nolen-Hoeksema, 2002), psychopathology (Twenge, Gentile, DeWall, Ma, Lacefield, & Schurtz, 2010), and narcissism (Twenge, Konrath, Foster, Campbell, & Bushman, 2008). Noguchi, Albarracín, Durantini, & Glasman (2007) examined interventions' recruitment and retention rates as factors that might relate to risk for acquiring or transmitting human immunodeficiency virus (HIV). Fischer, Hanke, and Sibling (2012) examined how social dominance orientation varies across 27 nations.

Standardizing Arithmetic Means across Studies.

If every study in a research literature operationalized the criterion of interest in exactly the same fashion, then metaanalyses can proceed without converting it to any other dimension (Bond, Wiitala, & Richard, 2003; Johnson & Boynton, 2008; Lipsey & Wilson, 2001). Doing so might be particularly advantageous when the measure is well known – measures of intelligence are good examples – as readers' familiarity with the measure helps make results easier to understand. Another alternative is mathematically converting results obtained on one scale to be equivalent with another scale. A mean value obtained on a 1-to-5 scale can be converted to the equivalent on a 1-to-7 scale or whatever target scale an analyst wishes to use across the literature of studies. Indeed, an argument can be made to move all such arithmetic means to their equivalents on a 0-to-100 scale, where 0 implies the lowest possible score and 100 is the maximum possible score. Targeting primary-level research, Cohen, Cohen, Aiken, & West (1999) advocated just such a procedure to convert means into percent of maximum

possible (POMP) scores:

$$M_{\text{POMP}} = \frac{M - \text{minimum possible score}}{\text{maximum possible score} - \text{minimum possible score}} \times 100, \quad (26.1)$$

where M is the observed mean. The advantage of the POMP procedure is that the transformed values now take a more immediately interpretable meaning – those close to 0 are low and those close to 100 are high, and 50 is the mid-point. Putting all observed M s in a literature on the POMP metric also serves the statistical purpose of putting the study results on a common metric. If effect sizes of association are the focus of the metaanalysis, now the POMP scores could serve as moderators of those T s. Lennon, Huedo-Medina, Gerwien, and Johnson (2012) provided an example of this moderator strategy, showing that HIV prevention interventions for women succeeded to a greater extent in samples for which depression (represented by POMP scores) was more marked.

Putting arithmetic means on the same metric also implies that they can plausibly be used as T s themselves. To date, this strategy has been relatively rare (for an example, see Fischer et al., 2012). In order to invoke this strategy, not only the arithmetic means must be put into POMP metric but also their accompanying standard deviations:

$$SD_{\text{POMP}} = \frac{SD}{\text{Maximum possible score} - \text{Minimum possible score}} \times 100. \quad (26.2)$$

As we explain in the next subsection, SD_{POMP} is needed to estimate the inverse variance that is used as a weight in analyses of T s.

Some cautions about POMP scores are in order. Converting study results to a common metric assumes that they can be scaled in this fashion. That is, values may not have the same meaning on every scale converted into a common metric (e.g., Rosenthal & Rosnow, 1991). Therefore, the same sample of individuals may exhibit varying levels on differing scales intended to measure the same

feature. If enough studies have multiple measures, metaanalyses can quantitatively test this assumption by examining whether different scales yield different M_{POMP} values.

Arithmetic Means versus Standardized Mean Difference Effect Sizes.

The fact that meta-analytic procedures allow use of the arithmetic mean as T might present a difficult decision for analysts who examine literatures in which two or more groups are compared on a continuous outcome (see [Table 26.1](#)). Historically, metaanalyses have defaulted to the standardized mean difference as T , but they could instead analyze the arithmetic means for each group. As Johnson and Boynton (2008) described, results from arithmetic means can provide even more detailed information about a literature than do results from the standardized mean difference. As we have noted, the latter form of T describes a difference between two means, where the sign of the T denotes whether one group is higher or lower than the other. Moderation patterns related to the standardized mean difference can leave unclear which of the two groups is changing most over the values of the moderator or moderators. As an example, Johnson and Boynton (2008) showed how mean sample age related positively to gender differences in social dominance orientation: As sample ages increased, standardized mean differences grew smaller. Yet, men may have decreased their support of social dominance, or women may have increased it. Johnson and Boynton used the arithmetic means separately for samples of females and males to show that the trend across the studies on the standardized mean difference index was primarily attributable to changes in the female samples. This example illustrates the use of *both* methods to gauge studies' effects.

There are some important caveats to using arithmetic means as T in a metaanalysis. First, many factors can affect the levels that arithmetic means take. For example, how positive participants are toward the position advocated in a persuasion experiment might be related to such factors as positive or negative mood, gender, personality traits, related attitudes, and of course the experimental condition itself. A metaanalysis could treat the mean for each condition as though it is an independent study, and if gender is the focus, subdivide each condition's data. Although some factors could be coded and used as moderators, many factors would not be possible to control. In contrast, metaanalyses that treat study information as two-variable effect sizes ([Table 26.1](#)) effectively control for the “noise” of variables that are not the focus of the metaanalysis. A

comparison between, say, males and females from the same study controls for every factor *except* gender (and its correlates). Second, no matter the scale used for standardization (including POMP), the inverse variance for the arithmetic mean, which is used for weighting in analyses, relies on each study's observed standard deviation (Lipsey & Wilson, 2001):

$$\text{Inverse variance} = \frac{n}{SD^2}. \quad (26.3)$$

One problem with POMP scores is related to the zero or near-zero standard deviations that may appear under some circumstances. For example, when observed arithmetic means take the maximum or the minimum possible value, their standard deviations will be zero, which implies that a weight cannot be calculated. Such studies might need to be omitted from analyses or examined with alternative assumptions.

Analyzing the Meta-Analytic Database

Preliminary Considerations.

The general steps involved in the analysis of any effect size, T , usually are the following: (1) aggregate effect sizes across the studies to determine the overall magnitude of the weighted mean T ; (2) analyze the consistency of the effect sizes across the studies; (3) diagnose statistical outliers among the effect sizes; (4) examine the distribution of effect sizes to determine whether any irregularities exist; and (5) perform tests of whether study attributes moderate the magnitude of the effect sizes.

Mean Effect Size and Homogeneity of Effect Sizes.

The model-testing procedures that we present are analogous to techniques used in data analysis in primary research and take advantage of weighted general linear models, where the weights are defined as the inverse variance, as we will explain. Models that divide results for categorical features are known as *subgroup analyses* or *categorical models*, and those that use continuous features are known as *meta-regressions* (which may also include categorical variables). Statistical analyses in metaanalysis differ from those in primary research in two main respects. The first difference pertains to the heterogeneity of the variances

ordinarily associated with the individual effect sizes, which would likely violate the homoscedasticity assumption of conventional regressions and ANOVAs (Hedges & Olkin, 1985), which is that standard deviations of the error terms do not vary and do not depend on predictors' values. Because this nonsystematic variance of an effect size is in general inversely proportional to the sample size of the study and sample sizes vary widely across the studies, the error variances of the effect sizes are ordinarily quite heterogeneous. Meta-analytic statistics aim to overcome this limitation (see the next subsection). The second difference between the statistical procedures of metaanalysis and primary research is that meta-analytic statistics permit an analysis of the consistency (or homogeneity) of the effect sizes across the studies – a highly informative analysis.

As a first step in a quantitative synthesis, the study outcomes are combined by averaging the T -values with each T_j for each study j is weighted by the reciprocal of its variance. The weighted mean effect size T_+ is a weighted average of the individual studies' effect sizes,

$$T_+ = \frac{\sum_{j=1}^k w_j T_j}{\sum_{j=1}^k w_j}, \quad (26.4)$$

where k is the number of effect sizes and w_j is the weight for each study j . The weights may be defined as a simple function of the sampling error associated with each effect size j , which follows *fixed-effects assumptions*. In this case, the inverse variance for each T serves as the weight (see examples in Table 26.2). Alternatively, analysts can define the weights to incorporate an estimate of the variance in the population of effect sizes, τ^2 (Hedges & Vevea, 1998), which follows *random-effects assumptions*. In either version of weighting, Equation 26.4 gives greater weight to the more reliably estimated study outcomes.

Cochran's (1954) Q evaluates the hypothesis that the effect sizes are homogeneous. Specifically, Q is a model specification statistic that evaluates how closely individual T_j correspond with T_+ ,

$$Q = \sum_{j=1}^k w_j (T_j - T_+)^2, \quad (26.5)$$

where k is the number of effect sizes in the class and W_j is based on fixed-effects assumptions (see examples in [Table 26.2](#)).⁷ Q has an approximate χ^2 distribution with $k - 1$ degrees of freedom. If Q is significant, the hypothesis of the homogeneity (or consistency) of the effect sizes is rejected, and heterogeneity is inferred. In other words, there is more variability in the observed T s than would be expected on the basis of the sampling error alone. In this event, the weighted mean effect size may not adequately describe the outcomes of the set of studies because it is likely that quite different mean effects exist in different groups of studies, and these differences may include differences in the direction (or sign) of the relation. In some subgroups of studies, X might have had a large positive effect on Y , and in other studies it might have had a smaller positive effect or even a negative effect on Y .

Values of Q are highly correlated with the numbers of T s entering into this statistic, making it difficult to compare levels of heterogeneity between metaanalyses and within portions of metaanalysis. To address this issue, Higgins and Thompson (2002) introduced a homogeneity index, I^2 , based on Q and its degrees of freedom. Values of I^2 range from 0 to 100%, where high values indicate more variability among the effect sizes and 0 implies homogeneity. Yet, I^2 is subject to the same conditions and qualifications as is Q (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006) such that both statistics are underpowered in small samples of studies. Moreover, values of I^2 at 25%, 50%, and 75% are often taken to describe small, moderate, and large amounts of heterogeneity, respectively. Yet, these cut points are best taken only as suggestions: Even a “small” I^2 can hide statistically significant variability in T s.

Even if a homogeneity test is nonsignificant, significant moderators could be present, especially when Q or I^2 are relatively large (Johnson & Turco, 1992). Also, Q and I^2 can be significant even though the effect sizes are very close in value, especially if the sample sizes are very large. Therefore, heterogeneity deserves careful interpretation, in conjunction with inspecting the values of the effect sizes. Nonetheless, in a metaanalysis that attempts to determine X 's impact on Y , rejecting the hypothesis of homogeneity could be troublesome because it implies that the association between these two variables likely is complicated by the presence of interacting conditions. Because analysts usually anticipate the presence of one or more moderators of effect-size magnitude, establishing that, overall, effect sizes lack homogeneity is ordinarily of no concern, unless analysts cannot determine the sources of the heterogeneity.

The fact that T_s may differ widely in magnitude should give analysts pause about the meaning of a weighted mean effect size, T_+ . In the face of heterogeneity, T_+ may lack a clear meaning, even if it is evaluated with random-effects assumptions, which are relatively conservative compared to fixed-effects assumptions. That is, incorporating random-effects assumptions will yield wider confidence intervals around T_+ than will those based on fixed-effects assumptions. Thus, a random-effects mean may disguise meaningful subpopulations of T_s .

In practice, the fixed-and random-effects variance components are summed to form new weights:

$$W_j = \frac{1}{\text{Variance}_{FE} + \tau^2}$$

where Variance_{FE} is the fixed-effects variance for each study and τ^2 is a constant for each study. The standard deviation of the population of effect sizes, τ , takes the same metric as T , and τ^2 is in the same metric as T^2 (for calculations, see Borenstein, Hedges, Higgins, & Rothstein, 2009). Using these weights in Equation 26.5 produces a mean based on random-effects assumptions. In the unlikely event that $\tau^2 = 0$, random-effects assumptions reduce to fixed-effects assumptions.

The variance, v_+ , of the weighted mean effect size T_+ is

$$v_+ = \frac{1}{\sum_{j=1}^k W_j}. \quad (26.6)$$

As a test for significance of this weighted mean effect size, one can calculate a confidence interval around this mean, based on its standard deviation, $T_+ \pm 1.96 \sqrt{v_+}$ where 1.96 is the unit-normal value for a 95% CI (assuming a nondirectional hypothesis). If the confidence interval (CI) includes zero (0.00), the value indicating exactly no difference, it may be concluded that, aggregated across all studies, there is no significant association between the independent and dependent variable (X and Y). The fixed-effects mean is known to be overpowered in the face of heterogeneity (Hedges & Vevea, 1998; Huedo-Medina, Sánchez Meca, & Marín Martínez, 2004). In other words, when study

results are inconsistent, a fixed-effects mean is more likely to reach statistical significance than is a random-effects mean, other factors being equal. Thus, assuming fixed-effects assumptions should be considered a relatively risky strategy of statistical inference.

Finally, analysts often present other measures of central tendency in addition to the weighted mean effect size (Borenstein et al., 2009). For example, the unweighted mean effect size shows the typical effect without weighting studies with larger sample sizes more heavily. A substantial difference in the values of the unweighted and weighted mean effect sizes suggests that one or more studies with large sample sizes may deviate from the rest of the sample. It is possible that larger studies used different methods than smaller studies did. Also, the median effect size describes a typical effect size but would be less affected than a mean effect size by outliers and other anomalies of the distribution of effect sizes.

Evaluating the Potential for Publication Bias.

Asymmetries in the distribution of effect sizes often are taken as evidence of publication bias, that is, the possibility that published results differ systematically from those that are not published (Sutton, 2009). *Funnel plots* (Light & Pillemer, 1984) are scatter plots of inverse variances versus effect sizes. When there is no publication bias, the scatterplot should take the shape of a funnel sitting on end in the sense that the effect sizes from smaller studies, which are less reliable, would show more scatter than the effect sizes from the larger studies, which would center on the best estimate of the population effect. Yet, if there is a publication bias in the literature, a funnel plot should reveal few entries in the smaller effect size portion of the graph for smaller sample sizes. There are many variations on such displays that are often quite sophisticated (Borman & Grigg, 2009). The most popular quantitative alternatives to examine for asymmetries include Egger, Smith, Schneider, and Minder's (1997) and Begg's (1985) tests, which provide estimates of the extent to which asymmetry is present in a distribution of effect sizes. Another popular tool is the trim-and-fill technique (Duval & Tweedie, 2000), which quantitatively assesses whether such asymmetries would change inferences about the significance of T_+ . An important caveat to all of these strategies is that each assumes a single population of effect sizes. Under heterogeneity, the tests may not be diagnostic of publication bias (e.g., Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006; Sutton, 2009).

Analysts sometimes calculate the number of studies averaging a null effect that would be necessary to bring an overall meta-analytic mean to the point of nonsignificance (Rosenthal, 1979). If this “fail-safe N ” (N_{fs}) is small, then the result seems less trustworthy. Specifically, one would calculate

$$N_{fs} = \frac{\left(\sum_{j=1}^k Z_j\right)^2}{z_{\alpha}^2}, \quad (26.7)$$

where k is the number of studies, Z_j is the unit normal value corresponding to a one-tailed test of significance, and z_{α} is the critical value (i.e., 1.645 for a one-tailed hypothesis). Orwin (1983) offered a variant of this equation that estimates N_{fs} directly from the mean weighted effect size. Although N_{fs} may have heuristic value in some instances, the equation for N_{fs} assumes that unretrieved studies would average null when in fact they may have the same pattern as the retrieved studies or even a reversed pattern. Also, it is difficult to evaluate the magnitude of N_{fs} because it has no statistical distribution theory (Becker, 2005).

Testing Models of Meta-Analytic Moderators.

To determine the relation between study characteristics and the magnitude of the effect sizes, analysts fit models using a form of weighted ordinary least squares regressions (for statistical methods, see Borenstein et al., 2009; Harbord & Higgins, 2008; Hedges & Olkin, 1985; Higgins & Thompson, 2004; Huedo-Medina & Johnson, 2010). Moderators, which are also called *effect modifiers*, can take the form of either categorical or continuous dimensions; they can be entered either solely (bivariate) or in a combined form. For example, in a continuous model, Hart, Albarracín, Eagly, Brechan, Lindberg, & Merrill (2009) found that, to the extent that information was more congenial, greater selective exposure resulted. Similarly, in a categorical model (also called subgroup analysis) they found that individuals preferred congenial over uncongenial information, especially when the issue was of high versus low value-relevance.

As noted, categorical and continuous features may be evaluated in meta-regression procedures, dummy-coding categorical variables as necessary. The unstandardized regression (b) coefficient(s) provide tests for the significance of the predictor's association with the effect sizes. Under fixed-effects assumptions,

the models use the inverse *variance* for each effect size as the weights. Such models are known to be overpowered in the face of heterogeneity (Hedges & Vevea, 1998). Under fixed-effects assumptions, the fit of meta-regression models is estimated by the error sum of squares statistic, Q_E , which has an approximate chi-square distribution with $k - p - 1$ degrees of freedom, where k is the number of effect sizes and p is the number of predictors (not including the intercept). Q_E can be converted to I^2 for evaluation.

Contemporary software permits easy incorporation of random-effects assumptions in such models. Such models are ordinarily *mixed-effects models* because differences between groups of T s (i.e., the slopes) are fixed and the constant (or intercept) follows random-effects assumptions (e.g., Harbord & Higgins, 2008). By convention, most analysts label these models *random-effects meta-regressions*, and this set of assumptions has become the most conventional for most meta-analytic situations. These models estimate the population variance, τ^2 , after removing the variance attributable to the moderators included in the model. Thus τ^2 can and does change from model to model. Commonly available output in these models includes I^2 residual, which is an assessment of the between-studies variability that is not explained by the model.

Outlier Diagnoses.

Because metaanalyses weight studies for their inverse variance, outliers with larger weights can dramatically alter meta-regression results (for a more general discussion of the topic of data outliers, see McClelland, Chapter 23 in this volume). Under such circumstances, these outliers can be removed from subsequent phases of the data analysis. Alternatively, T s that are far distant from other T s can be winsorized so that they are not so extreme. The same can be done for inverse variance estimates that are relatively extreme. Outliers might be detected in many ways, but one that is highly recommended is to examine the residuals in meta-regression models.

Depictions of Effect Size Magnitude.

In some instances, visual presentations can assist greatly in the interpretation of meta-analytic results (Borman & Grigg, 2009; Johnson & Huedo-Medina, 2011). For example, visually examining study outcomes enhances the analyst's potential for finding anomalies in the meta-analytic data. By examining how effect sizes vary over the range of a moderator, an analyst may determine that

effect sizes are related to a continuous predictor in a nonmonotonic fashion – an outcome that would not be detected by the linear regressions that have been described to this point in the chapter. Meta-regression models may include tests of nonlinear associations, yet unless nonmonotonic associations are expected on an a priori basis, they are unlikely to be discovered except by the use of visual displays.

Depictions of model results in either graphical or tabled form can help describe results in presentations and written reports. Johnson and Huedo-Medina (2011) described *the moving constant technique*, with which analysts can use meta-regression to create graphs of effect sizes plotted against moderator values, including confidence bands around the meta-regression line. This technique can also be used to estimate mean effect size values and confidence intervals at moderator values of interest. Specifically, analysts may move the intercept to reflect interesting points along or beyond a range of independent variable values. For example, Lennon *et al.* (2012) found that HIV prevention efforts for women succeeded better for samples with higher baseline depression. Using the moving constant technique, they estimated the amount of risk reduction for samples with the highest mean levels of depression to be large and significant, whereas for samples with lower levels of depression, on average, interventions failed to impact risk (see Figure 26.1). Results presented in this form help show for what levels of a moderator an effect exists. Such estimates, in turn, can be highly informative when interpreting the nature of the phenomenon being studied in the metaanalysis, especially when a comparison to an absolute or a practical criterion is important. The moving constant technique also permits analysts to estimate confidence intervals for an effect size at particular values of one or more independent variables (and thus to avoid artificially dichotomizing continuous predictor variables).

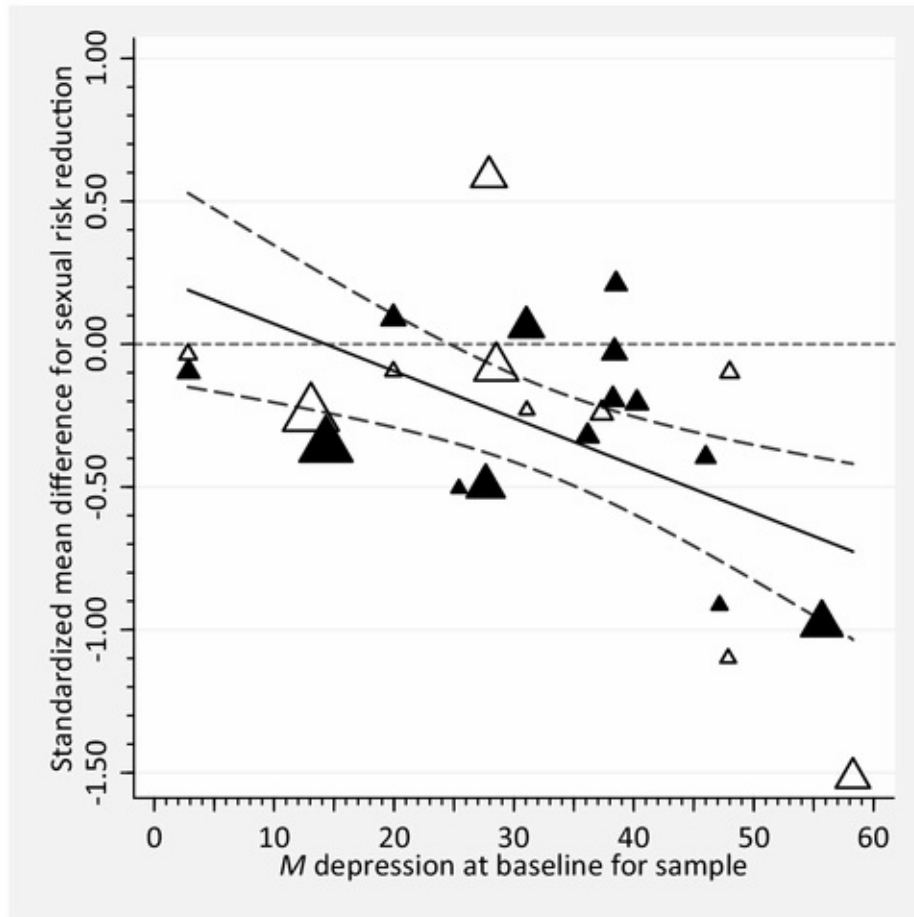


Figure 26.1. Sexual risk reduction following a behavioral intervention as a function of each sample's baseline depression. Sexual risk behavior declined following the intervention at the last available follow-up to the extent that samples had higher levels of baseline depression (treatment [control] group effects appear as darker [white] triangles and the size of each plotted value reflects its weight in the analysis). The solid regression line indicates trends across initial levels of depression; dashed lines provide 95% confidence bands for these trends. Reproduced from Lennon *et al.* (2012).

Dealing with Nonindependent Effect Sizes.

We have indicated that, as a general rule, it is wise to represent studies' participants only once in effect size calculations. Thus, analysts should ordinarily combine effect sizes representing conceptually similar measures from any given study. If such effect sizes were not combined, the nonindependence that would result could have several effects on the findings of a metaanalysis, depending on the source of the nonindependence (Gleser & Olkin, 2009). If the nonindependence results from producing more than one effect size from the

same participants on correlated measures, the metaanalysis will be likely to reach a liberal estimate of the significance of the weighted mean effect size: Its *CI* will grow tighter. Including more effect sizes from the same groups of participants may also affect inferences from model-fit statistics (Q or I^2).

Despite these concerns, representing studies multiple times may be defensible to address certain meta-analytic questions. One such question is whether an effect generalizes across various types of measures of a dependent variable. In such a case, the analyst could examine a model to determine if the effect sizes differed according to the type of measure used. If the synthesis forgoes this analysis to uphold the assumption that effect sizes are independent, potentially valuable information about a moderator would be lost. Therefore, one defensible strategy is to conduct a two-stage metaanalysis that shifts its units of analysis (Cooper, 2010). In the first stage, the metaanalysis would address the study-level effect sizes, which represent the information from each study only once. A second stage would divide study outcomes into the various groupings specified by moderators and would permit information for a group of study participants to appear more than once, in order to examine the differences across the moderator (for examples of this strategy, see Gerrard, Gibbons, & Bushman, 1996; Kolodziej & Johnson, 1996). This ordering of the stages enables analysts to learn the overall, more general pattern in the literature prior to answering specific questions about moderators. This combination of approaches should help allay concerns about nonindependence while still yielding the desired information. Other alternatives include (a) using multivariate procedures for the analysis of multiple effect sizes from each study (Gleser & Olkin, 2009); (b) representing effect sizes nested within studies in terms of multilevel models (Hedges, 2009); or (c) pursuing individual-level metaanalyses of studies whose raw data are available, in a practice also known as integrated data analysis (Cooper & Patall, 2009; Stewart, Tierney, & Burdett, 2005). This latter option is often considered the gold standard of metaanalysis when the individual-level studies reviewed are highly representative of the often much larger literatures for which only study-level effects are available.

Interpretations of Effect Size Indexes of Association.

Cohen (1969, 1988) tentatively proposed some guidelines for judging effect magnitude, based on his informal analysis of the magnitude of effects commonly yielded by psychological research. Cohen intended “that medium represents an effect of a size likely to be visible to the naked eye of a careful observer”

(Cohen, 1992, p. 156). He intended that small effect sizes be “noticeably smaller yet not trivial,” and that large effect sizes “be the same distance above medium as small is below it” (p. 156). As Table 26.3 shows, a “medium” effect turned out to be about $d = 0.50$ and $r = .30$, equivalent to the difference in intelligence scores between clerical and semiskilled workers. A “small” effect size was about $d = 0.20$ and $r = .10$, equivalent to the difference in height between 15-and 16-year-old girls. Finally, a large effect was about $d = 0.80$ and $r = .50$, equivalent to the difference in intelligence scores between college professors and college freshmen. Although these impressionistic guidelines for magnitude of effects are frequently cited, there are caveats about particular effect size indexes’ magnitude (McGrath & Meyer, 2006). Many alternatives exist for interpreting the magnitude of effects.

Table 26.3. Cohen's (1969) Guidelines for Magnitude of d and r

	Effect Size Metric		
Size	d	R	r^2
Small	0.20	.10	.01
Medium	0.50	.30	.09
Large	0.80	.50	.25

One popular way to interpret mean effect sizes is to derive the equivalent r and square it. This procedure shows how much variability would be explained by an effect of the magnitude of the mean effect size (see Table 26.3). Thus, a mean d of 0.50 produces an R^2 of .09. However, this value must be interpreted carefully because R^2 , or variance explained, is a directionless effect size. Therefore, if the individual effect sizes that produced the mean effect size varied in their signs (i.e., the effect sizes were not all negative or all positive), the variance in Y explained by the predictor X , calculated for each study and averaged, would be larger than this simple transformation of the mean effect size.

A number of methodologists have argued that even quantitatively small

effects can be quite consequential (e.g., Abelson, 1985; Prentice & Miller, 1992; Rosenthal, 1990; Ross & Nisbett, 1991), and some have provided tools to help show how meaningful an implied effect size is in application. These tools include Rosenthal and Rubin's (1982) binomial effect size display (for caveats, see Thompson & Schumacker, 1997), McGraw and Wong's (1992) common language effect size statistic index, and Rosenthal and Rubin's (1994) counternull statistic. In using such tools, the meta-analyst attempts to reach some conclusion about how much the effect matters in terms of some tangible outcome.

Another method of interpreting the magnitude of effect sizes is to compare them with effect sizes in similar domains in which magnitude is already known. For example, Eagly (1995) argued that claims that sex-related differences in behavior are necessarily small should be evaluated in relation to the magnitude of other known effects in psychology. Following this strategy, Bettencourt and Miller (1996) compared the magnitude of sex-related differences in aggression to the magnitude of the effect of provocation on aggression, which was derived from the same sample of studies. More generally, meta-analysts ought to compare the magnitude of a newly derived meta-analytic effect size to the magnitude of known effects in the same or related research areas. It is also important to consider the implications of effect sizes in metrics that are sensible in natural settings (e.g., number of lives saved by treatments, proportions of girls and boys admitted to selective educational programs, given a particular ability sex difference).

Many aspects of studies' methods can constrain effect magnitude. As we noted in the section on Correcting Effect Sizes for Bias, effects are larger or smaller depending on factors such as reliability of measures, heterogeneity of the participant population, and so on. Some of these factors lend themselves to bias corrections, and a study's effect size depends on whether corrections have been applied for such problems. In addition, characteristics of the situation in which experiments are carried out can increase or reduce the impact that experimental manipulations and individual-difference variables have on dependent variables (Prentice & Miller, 1992). Analysts should code studies for the presence of a wide range of such factors, to account for effect size variance produced by studies' nonequivalence on such factors.

Conducting and Evaluating MetaAnalyses

Our treatment of meta-analytic methods has stressed the importance of high standards in conducting and evaluating these reviews. From the preceding sections of this chapter, a picture of a high-quality metaanalysis emerges:

1. Define the research problem clearly and, if possible, define hypotheses prior to commencing with the metaanalysis.
2. Use highly inclusive search strategies that locate unpublished as well as published studies.
3. Be explicit in the criteria for selecting studies and, if possible, define these a priori.
4. Thoroughly and accurately code moderator variables and other study-relevant information.
5. Represent study outcomes with high accuracy.
6. Conduct meta-analytic models, maintaining fidelity to the statistics' assumptions.
7. Interpret findings carefully in relation to the assumptions that underlie both individual studies and the metaanalysis itself.

Each of these dimensions appears in Shea et al.'s (2007) recent quality-coding protocol for metaanalysis. Nonetheless, even a quantitative review that meets high standards does not necessarily constitute an important scientific contribution.

One factor affecting the scientific contribution of a synthesis is that its conclusions are limited by the quality of the data that are synthesized. Serious methodological faults that are endemic in a research literature may well handicap a synthesis, unless it is designed to shed light on the influence of these faults. Also, to be regarded as important, the review must address an interesting question. Similarly, unless the paper reporting a metaanalysis “tells a good story,” its full value may go unappreciated by readers. Although there are many paths to a good story, Sternberg's (1991) recommendations to authors of reviews are instructive: pick interesting questions, challenge conventional understandings if at all possible, take a unified perspective on the phenomenon, offer a clear take-home message, and write well.

Some reports of research syntheses may fail to tell a good story because they are overly complex. This complexity may arise from the fact that quantitative synthesis forces the reviewer to study the minute details of the studies' methods and findings. Although this close scrutiny can yield valuable insights, it may also

foster a review that reflects too many complexities and thereby obscures its major findings. In short, even if a synthesis happens to solve a time-honored problem, it will have a poor reception if its message is mired in a forest of distracting minutiae. Excellent organization and skillful writing can overcome this challenge.

Although many critiques of metaanalyses have taken a narrative form by discussing their methods and findings, the most informative critiques take a quantitative approach by empirically evaluating the findings and conclusions. A critique that may seem reasonable based on sheer logic may become overwhelming when supported by appropriate data. In this manner, scientific disputes can be arbitrated by empirical tests. In primary research, the most influential critiques take the form of replications with variations, often showing how an effect disappears once a confound is controlled. Similarly, criticism of quantitative syntheses proceeds most effectively in an empirical fashion. In our view, replications of meta-analytic reviews should become more frequent, so that faults that may be present in one review are evaluated or eliminated in later reviews.

With metaanalyses having become commonplace, investigators should anticipate the recycling of their findings in metaanalyses. They should therefore redouble their efforts to report the method and results of their studies as accurately and completely as possible, aided by supplements and archives. Researchers can find excellent guidance in the Journal Article Reporting Standards (JARS) presented in the *Publication Manual* of the American Psychological Association (2010). In particular, for experimental studies, a table of means and standard deviations for each primary dependent variable, reported for all cells of the design, should be conventional. It is very helpful if exact statistics are provided even for auxiliary effects that may be nonsignificant (e.g., the comparison of female and male participants). For correlational studies, a complete matrix of the variables' intercorrelations should be conventional.

Additional Resources on Research Synthesis

Hunt (1997) provides a compelling and highly readable history on research synthesis. Essential reference works for conducting metaanalyses are *The Handbook of Research Synthesis and MetaAnalysis*, edited by Cooper, Hedges, and Valentine (2009), as well as texts by Borenstein *et al.* (2009), Card (2012), Cooper (2010), Hedges and Olkin (1985), and Lipsey and Wilson (2001). Two

of these offer either commercial software (Borenstein et al., 2009) or open-access macros for popular statistical platforms (Lipsey & Wilson, 2001). Viechtbauer (2010) authored a flexible and powerful set of tools for the open-source statistics software package, R. Other works may be particularly valuable for other aspects of metaanalysis: Hunter and Schmidt (2004) extensively addressed corrections to effect sizes; Glass et al.'s (1981) book remains a good source on derivations of effect sizes.

The Future of MetaAnalysis in Social and Personality Psychology

The growing numbers of studies on personality and social psychology's central phenomena dictate that, in the future, greater importance will be accorded to high-quality metaanalyses of these knowledge bases. In our opinion, the quality of metaanalyses has improved over the past decades. Metaanalysis should foster a healthy interaction between primary research and research synthesis, at once summarizing old research and suggesting promising directions for new research. One misperception that psychologists sometimes express is that a metaanalysis represents a point beyond which nothing more needs to be known. On the contrary, carefully conducted metaanalyses can often be the best medicine for a literature, by documenting the robustness with which certain associations are attained, resulting in a sturdier foundation on which future theories may rest. In addition, metaanalyses can show where knowledge is at its thinnest, to help plan additional, primary-level research (Wood & Eagly, 2009). As a consequence of a carefully conducted metaanalysis, new studies can be designed with the complete existing literature in mind and therefore have a better chance of contributing new knowledge. In this fashion, scientific resources can be directed more efficiently toward gains in knowledge.

The advent of computerized and readily accessible databases of psychological research literatures (e.g., PsycINFO) has meant that less time and financial resources are necessary to conduct metaanalyses than in the past. Despite these gains, psychologists face severe limitations in obtaining access to the data underlying completed research. In contrast to some other scientific fields (e.g., sociology, political science), few raw data from primary research are archived in psychology, and this omission greatly limits the opportunity for reviewers to perform the secondary analyses that can produce effect sizes for phenomena that have not been adequately reported. Primary researchers are often unable or

unwilling to provide needed statistical information when they are contacted directly. Routine data archiving in a central location would remedy this unfortunate situation (Cooper et al., 1997).

Psychologists and other scientists rely more and more on metaanalyses to inform them about the knowledge that has accumulated in their research. Although metaanalysis might become the purview of an elite class of researchers who specialize in research integration, as Schmidt (1992) argued, we believe that, on the contrary, metaanalysis will become a routine part of graduate training in many fields. With computer programs to aid calculations, most researchers should be able to integrate findings across studies as a normal and routine part of their research activities. Indeed, the publication trends⁸ within social and personality psychology that we portray in Figure 26.2 suggest that this phenomenon is occurring. Metaanalysis has become central to these areas of research and to many others.

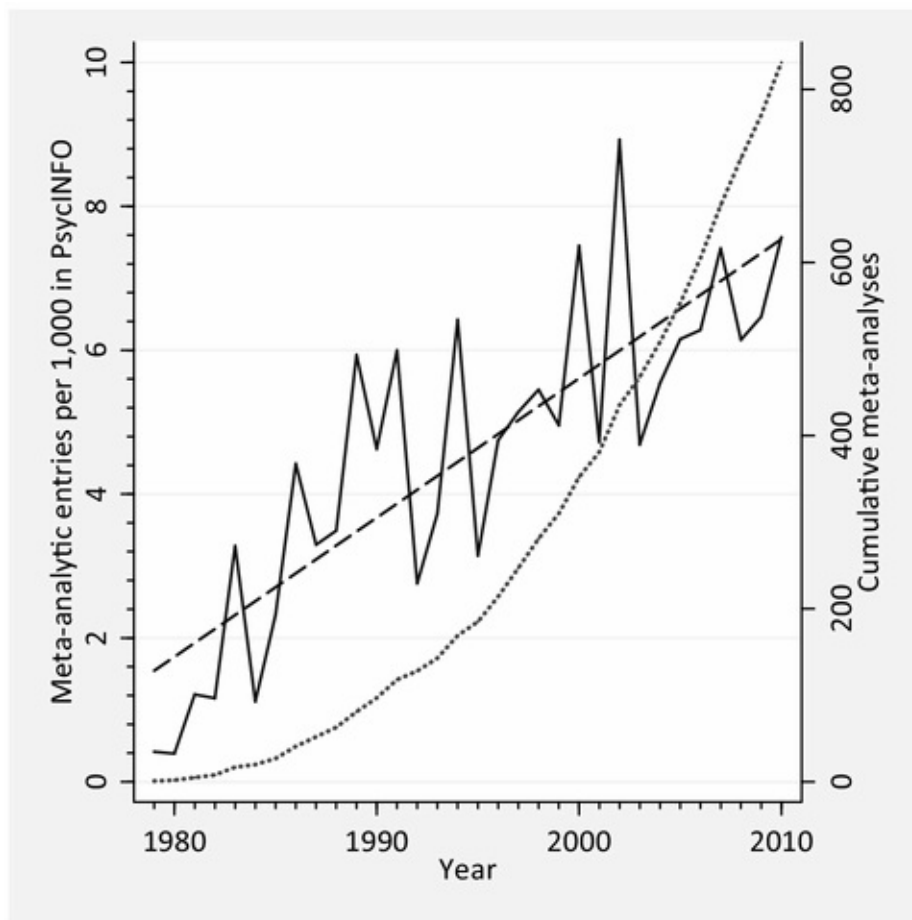


Figure 26.2. Publication trends in metaanalyses in social and personality psychology, where the solid line plots the number of reports per year per 1,000 recorded in PsycINFO; the dashed line is the best-fitting linear trend (both on the

left axis), and the dotted line represents cumulative meta-analytic reports (right axis).

References

- Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, 97, 129–133.
- Albarracín, D., Johnson, B. T., Fishbein, M., & Muellerleile, P. A. (2001). Theories of reasoned action and planned behavior as models of condom use: A metaanalysis. *Psychological Bulletin*, 127(1), 142–161.
- American National Election Studies. (2012). Databases related to political trends. Retrieved from <http://www.electionstudies.org/>
- American Psychological Association. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, 63(9), 839–851.
- American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.
- Baas, M., De Dreu, C. K. W., & Nijstad, B. A. (2008). A metaanalysis of 25 years of mood-creativity research: Hedonic tone, activation, or regulatory focus? *Psychological Bulletin*, 134(6), 779–806.
- Balliet, D., Li, N. P., Macfarlan, S. J., & Van Vugt, M. (2011). Sex differences in cooperation: A meta-analytic review of social dilemmas. *Psychological Bulletin*, 137(6), 881–909.
- Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, 41, 257–278.
- Becker, B. J. (2005). Failsafe N or file-drawer number. In H. Rothstein, A. Sutton, & M. Borenstein (Eds.), *Publication bias in metaanalysis: Prevention, assessment and adjustments* (pp. 111–126). Chichester, UK: Wiley.
- Begg, C. B. (1985). A measure to aid in the interpretation of published clinical trials. *Statistics in Medicine*, 4, 1–9.
- Bettencourt, B. A., & Miller, N. (1996). Gender differences in aggression as a function of provocation: A metaanalysis. *Psychological Bulletin*, 119, 422–

447.

- Bond, C. F., & Titus, L. J. (1983). Social facilitation: A metaanalysis of 241 studies. *Psychological Bulletin*, 94, 265–292.
- Bond, C. F., Wiitala, W. L., & Richard, F. D. (2003). Metaanalysis of raw mean differences. *Psychological Methods*, 8, 406–418.
- Bond, R., & Smith, P. B. (1996). Culture and conformity: A metaanalysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin*, 119, 111–137.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to metaanalysis*. Chichester, UK: Wiley.
- Borman, G. D., & Grigg, J. A. (2009). Visual and narrative interpretation. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and metaanalysis* (2nd ed., pp. 497–519). New York: Russell Sage Foundation.
- Brubaker, T. H., & Powers, E. A. (1976). The stereotype of “old”: A review and alternative approach. *Journal of Gerontology*, 31, 441–447.
- Bushman, B. J., & Wang, M. C. (1996). A procedure for combining sample standardized mean differences and vote counts to estimate the population standardized mean difference in fixed effects models. *Psychological Methods*, 1, 66–80.
- Cafri, G., Kromrey, J. D., & Brannick, M. T. (2009). A SAS macro for statistical power calculations in metaanalysis. *Behavior Research Methods*, 41(1), 35–46.
- Campbell, D. T., & Stanley, J. T. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.
- Card, N. A. (2012). *Applied metaanalysis for social science research*. New York: Guilford Press.
- Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, 26, 347–372.
- Chiaburu, D. S., Oh, I. S., Berry, C. M., Li, N., & Gardner, R. G. (2011). The five-factor model of personality traits and organizational citizenship behaviors: A metaanalysis. *Journal of Applied Psychology*, 96(6), 1140–1166.

- Cingranelli-Richards Human Rights Project. (2012). Human rights data across nations. Retrieved April 6, 2012, from <http://www.humanrightsdata.org/>
- Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, 10(1), 101–129.
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstances for POMP. *Multivariate Behavioral Research*, 34, 315–346.
- Cooper, H. (1979). Statistically combining independent studies: Metaanalysis of sex differences in conformity research. *Journal of Personality and Social Psychology*, 37, 131–146.
- Cooper, H. (2010). *Research synthesis and metaanalysis: A step-by-step approach* (4th ed.). Los Angeles: Sage.
- Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods*, 2(4), 447–452.
- Cooper, H., & Hedges, L. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.
- Cooper, H., Hedges, L., & Valentine, J. (Eds.). (2009). *The handbook of research synthesis and metaanalysis* (2nd ed.). New York: Russell Sage Foundation.
- Cooper, H., & Patall, E. A. (2009). The relative benefits of metaanalysis conducted with individual participant data versus aggregated data. *Psychological Methods*, 14, 165–176.
- Cooper, H., & Rosenthal, R. (1980). Statistical versus traditional procedures for

- summarizing research findings. *Psychological Bulletin*, 87, 442–149.
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: HarperCollins.
- Darlington, R. B., & Hayes, A. F. (2000). Combining independent p values: Extensions of the Stouffer and binomial methods. *Psychological Methods*, 5(4), 496–515.
- Deeks, J. J., Dinnes, J., D’Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., Petticrew, M. & Altman, D. J. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment*, 7(27), 1–179.
- Dickersin, K. (1997). How important is publication bias? A synthesis of available data. *AIDS Education and Prevention*, 9 (Suppl. A), 15–21.
- Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Metaanalysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, 1, 170–177.
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in metaanalysis. *Biometrics*, 56, 455–463.
- Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. Hillsdale, NJ: Erlbaum.
- Eagly, A. H. (1995). The science and politics of comparing women and men. *American Psychologist*, 50, 145–158.
- Eagly, A. H., & Crowley, M. (1986). Gender and helping behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin*, 100, 283–308.
- Eagly, A. H., Johannesen-Schmidt, M. C., & van Engen, M. L. (2003). Transformational, transactional, and laissez-faire leadership styles: A metaanalysis comparing women and men. *Psychological Bulletin*, 129(4), 569–591.
- Eagly, A. H., Karau, S. J., & Makhijani, M. G. (1995). Gender and the effectiveness of leaders: A metaanalysis. *Psychological Bulletin*, 117(1), 125–145.
- Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. (1992). Gender and the

- evaluation of leaders: A metaanalysis. *Psychological Bulletin*, 111, 3–22.
- Eagly, A. H., & Steffen, V. J. (1986). Gender and aggressive behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin*, 100, 309–330.
- Edwards, A. W. F. (1963). The measure of association in a 2×2 table. *Journal of the Royal Statistical Society: Series A*, 126, 109–114.
- Edwards, J. H. (1957). A note on the practical interpretation of 2×2 tables. *British Journal of Preventive & Social Medicine*, 11, 73–78.
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in metaanalysis detected by a simple graphical test. *British Medical Journal*, 315, 629–634.
- Ernst, C., & Angst, J. (1983). *Birth order: Its influence on personality*. New York: Springer-Verlag.
- Feingold, A. (1995). The additive effects of differences in central tendency and variability are important in comparisons between groups. *American Psychologist*, 50, 5–13.
- Fischer, R., & Boer, D. (2011). What is more important for national well-being: Money or autonomy? A metaanalysis of well-being, burnout, and anxiety across 63 societies. *Journal of Personality and Social Psychology*, 101(1), 164–184.
- Fischer, R., Hanke, K., & Sibley, C. G. (2012). Cultural and institutional determinants of Social Dominance Orientation: A cross-cultural metaanalysis of 27 societies. *Political Psychology*, 33, 437–467.
- Fisher, R. A. (1921). On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron*, 1, 1–32.
- Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, 19, 151–156.
- Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A metaanalysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin*, 138(2), 296–321.
- Gapminder (2012). Social, political, and health databases. Retrieved April 6,

2012, from <http://www.gapminder.org/>

General Social Survey (2012). Survey data regarding the U.S. population. Retrieved April 6, 2012, from <http://www3.norc.umd.edu/gss/>

Gerrard, M., Gibbons, F. X., & Bushman, B. J. (1996). Relation between perceived vulnerability to HIV and precautionary sexual behavior. *Psychological Bulletin*, 119(3), 390–409.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Metaanalysis in social research*. Beverly Hills, CA: Sage.

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and metaanalysis* (2nd ed., pp. 357–376). New York: Russell Sage Foundation.

Goldberg, W. A., Prause, J., Lucas-Thompson, R., & Himself, A. (2008). Maternal employment and children's achievement in context: A metaanalysis of four decades of research. *Psychological Bulletin*, 134, 77–108.

Green, S. K. (1981). Attitudes and perceptions about the elderly: Current and future perspectives. *International Journal of Aging and Human Development*, 13, 99–119.

Greenhouse, J. B., & Iyengar, S. (2009). Sensitivity analysis and diagnostics. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and metaanalysis* (2nd ed., pp. 417–433). New York: Russell Sage Foundation.

Greenwald, S., & Russell, R. L. (1991). Assessing rationales for inclusiveness in meta-analytic samples. *Psychotherapy Research*, 1(1), 17–24.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A metaanalysis. *Psychological Bulletin*, 136, 495–525.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin*, 85, 845–857.

Harbord, R. M., & Higgins, J. P. T. (2008). Meta-regression in Stata. *Stata Journal*, 8, 493–519.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill,

- L. (2009). Feeling validated versus being correct: A metaanalysis of selective exposure to information. *Psychological Bulletin*, 135(4), 555–588.
- Hays, W. L. (1988). *Statistics* (4th ed.). Fort Worth, TX: Harcourt Brace Jovanovich.
- Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107–128.
- Hedges, L. V. (1990). Directions for future methodology. In K. W. Wachter & M. L. Straf (Eds.), *The future of metaanalysis* (pp. 11–26). New York: Russell Sage Foundation.
- Hedges, L.V. (2009). Statistical considerations. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and metaanalysis* (2nd ed., pp. 37–46). New York: Russell Sage Foundation.
- Hedges, L. V., & Becker, B. J. (1986). Statistical methods in the metaanalysis of research on gender differences. In J. S. Hyde & M. C. Linn (Eds.), *The psychology of gender: Advances through metaanalysis* (pp. 14–50). Baltimore, MD: Johns Hopkins University Press.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for metaanalysis*. Orlando, FL: Academic Press.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in metaanalysis. *Psychological Methods*, 3, 486–504.
- Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods*, 1, 154–169.
- Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a metaanalysis. *Statistics in Medicine*, 21, 1539–1558.
- Higgins, J. P. T., & Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics & Medicine*, 23, 1663–1682.
- Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.
- Huedo-Medina, T. B., & Johnson, B. T. (2010). *Modelos estadísticos en meta-análisis* [Statistical models in metaanalysis]. Series in Methodology and Data

Analysis in Social Sciences. La Coruña, Spain: Netbiblio.

- Huedo-Medina, T. B., Sánchez-Meca J., & Marín-Martínez F. (2004). Estimación del tamaño del efecto medio en un meta-análisis: Una comparación entre los modelos de efectos fijos y aleatorios. [Estimating the average of the effect size in a metaanalysis: A comparison between fixed-and random-effects models.] *Metodología de las Ciencias del Comportamiento*, Volumen Especial, 307–315.
- Huedo-Medina, T. B., Sánchez-Meca J., Marín-Martínez F., & Botella J. (2006). Assessing heterogeneity in metaanalysis: Q statistic or I^2 index? *Psychological Methods*, 11, 193–206.
- Hunt, M. (1997). *How science takes stock: The story of metaanalysis*. New York: Russell Sage Foundation.
- Hunter, J. E., & Schmidt, F. L. (1990). *Methods of metaanalysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of metaanalysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: Sage Publication.
- International Labor Organization. (2012). Databases. Retrieved April 6, 2012, from <http://www.ilo.org/global/lang-en/index.htm>.
- International Social Survey Programme. (2012). Survey data from many nations. Retrieved April 6, 2012, from <http://www.issp.org/>
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. doi:10.1371/journal.pmed.0020124.
- Johnson, B. T. (1993). *DSTAT 1.10: Software for the meta-analytic review of research literatures*. Hillsdale, NJ: Erlbaum.
- Johnson, B. T., & Boynton, M. H. (2008). Cumulating evidence about the social animal: Metaanalysis in social-personality psychology. *Social and Personality Psychology Compass*, 2(2), 817–841.
- Johnson, B. T., & Eagly, A. H. (1989). Effects of involvement on persuasion: A metaanalysis. *Psychological Bulletin*, 106(2), 290–314.
- Johnson, B. T., & Huedo-Medina, T. B. (2011). Depicting estimates using the intercept in meta-regression models: The moving constant technique.

Research Synthesis Methods, 2(3), 204–220.

- Johnson, B. T., Scott-Sheldon, L. A., & Carey, M. P. (2010). Meta-synthesis of health behavior change metaanalyses. *American Journal of Public Health*, 100(11), 2193–2198.
- Johnson, B. T., & Turco, R. (1992). The value of goodness-of-fit indices in metaanalysis: A comment on Hall and Rosenthal. *Communication Monographs*, 59, 388–396.
- Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., & Johnson, B. T. (2008). Initial severity and antidepressant benefits: A metaanalysis of data submitted to the FDA. *PLoS Medicine*, 5, 260–268.
- Kite, M. E., & Johnson, B. T. (1988). Attitudes toward the elderly: A metaanalysis. *Psychology and Aging*, 3, 233–244.
- Kite, M. E., & Whitley, B. R. (1996). Sex differences in attitudes toward homosexual persons, behaviors, and civil rights: A metaanalysis. *Personality and Social Psychology Bulletin*, 22(4), 336–353.
- Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. (2011). Are leader stereotypes masculine? A metaanalysis of three research paradigms. *Psychological Bulletin*, 137, 616–642.
- Kolodziej, M. E., & Johnson, B. T. (1996). Effects of interpersonal contact on acceptance of individuals diagnosed with mentally illness: A research synthesis. *Journal of Consulting and Clinical Psychology*, 64, 1387–1396.
- Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking “big” personality traits to anxiety, depressive, and substance use disorders: A metaanalysis. *Psychological Bulletin*, 136(5), 768–821.
- Lau, J., Ioannidis, J. P. A., Terrin, N., Schmid, C. H., Olkin, I. (2006). The case of the misleading funnel plot. *British Medical Journal*, 333, 597–600.
- Lee, I., Pratto, F., & Johnson, B. T. (2011). Intergroup consensus/disagreement in support of group-based hierarchy: An examination of socio-structural and psycho-cultural factors. *Psychological Bulletin*, 137(6), 1029–1064.
- Lennon, C. A., Huedo-Medina, T. B., Gerwien, D. P., & Johnson, B. T. (2012). A role for depression in sexual risk reduction for women? A metaanalysis of HIV prevention trials with depression outcomes. *Social Science & Medicine*,

75(4), 688--698.

- Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.
- Lipsey, M. W. (2009). Identifying potentially interesting variables and analysis opportunities. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and metaanalysis* (2nd ed., pp. 148–158). New York: Russell Sage Foundation.
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from metaanalysis. *American Psychologist*, 48(12), 1181–1209.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical metaanalysis*. Thousand Oakes, CA: Sage.
- Lutsky, N. (1981). Attitudes toward old age and elderly persons. In C. Eisdorfer (Ed.), *Annual review of gerontology and geriatrics* (Vol. 1, pp. 287–336). New York: Springer.
- Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) metaanalysis. *Psychological Bulletin*, 132(6), 895–919.
- Marcus-Newhall, A., Pedersen, W. C., Carlson, M., & Miller, N. (2000). Displaced aggression is alive and well: A meta-analytic review. *Journal of Personality and Social Psychology*, 78(4), 670–689.
- McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d . *Psychological Methods*, 11(4), 386–401.
- McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111, 361–365.
- Miller, N., Lee, J., & Carlson, M. (1991). The validity of inferential judgments when used in theory-testing metaanalysis. *Personality and Social Psychology Bulletin*, 17(3), 335–343.
- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and metaanalyses: The PRISMA statement. *Journal of Clinical Epidemiology*, 62(10), 1006–1012.
- Morris, S. B., & DeShon, R. P. (1997). Correcting effect sizes computed from factorial analysis of variance for use in metaanalysis. *Psychological Methods*,

2, 192–199.

- Morris, S. B., & DeShon, R. P. (2002) Combining effect size estimates in metaanalysis with repeated measures and independent groups by designs. *Psychological Methods*, 7, 105–125.
- Morrison, D. F. (1976). *Multivariate statistical methods*. New York: McGraw-Hill.
- Moyer, A., & Finney, J. W. (2002). Randomized versus nonrandomized studies of alcohol treatment: Participants, methodological features and posttreatment functioning. *Journal of Studies on Alcohol*, 63(5), 542–550.
- Mullen, B., & Felleman, V. (1989). Tripling in the dorns: A meta-analytic integration. *Basic and Applied Social Psychology*, 11, 33–43.
- Myers, J. L., & Well, A. D. (1991). *Research design and statistical analysis*. New York: Harper Collins.
- Noguchi, K., Albarracín, D., Durantini, M. R., & Glasman, L. R. (2007). Who participates in which health promotion programs? A metaanalysis of motivations underlying enrollment and retention in HIV-prevention interventions. *Psychological Bulletin*, 133, 955–975.
- Nouri, H., & Greenberg, R. H. (1995). Meta-analytic procedures for estimation of effect sizes in experiments using complex analysis of variance. *Journal of Management*, 21, 801–812.
- Orwin, R. G. (1983). A fail-safe *N* for effect size in metaanalysis. *Journal of Educational Statistics*, 8, 157–159.
- Orwin, R. G., & Vevea, J. L. (2009). Evaluating coding decisions. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and metaanalysis* (2nd ed., pp. 177–203). New York: Russell Sage Foundation.
- Pearson, K. (1895). Contributions to the mathematical theory of evolution II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London*, 186, 343–414.
- Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, 90(5), 751–783.
- Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive.

Psychological Bulletin, 112(1), 160–164.

Ray, J. W., & Shadish, W. R. (1996). How interchangeable are different estimators of effect size? *Journal of Consulting and Clinical Psychology*, 64, 1316–1325.

Reed, J. G., & Baxter, P. M. (2009). Using reference databases. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and metaanalysis* (2nd ed., pp. 73–101). New York: Russell Sage Foundation.

Rhodes, N., & Wood, W. (1992). Self-esteem and intelligence affect influenceability: The mediating role of message reception. *Psychological Bulletin*, 111, 156–171.

Rosenthal, R. (1968). Experimenter expectancy and the reassuring nature of the null hypothesis decision procedure. *Psychological Bulletin*, 70 (6, Pt. 2), 30–47.

Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, 86, 638–641.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist*, 45, 775–777.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Beverly Hills, CA: Sage.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. New York: Cambridge University Press.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Rosenthal, R., & Rubin, D. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 3, 377–415.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74, 166–169.

- Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science*, 5, 329–334.
- Ross, L., & Nisbett, R. (1991). *The person and the situation: Perspectives of social psychology*. New York: McGraw-Hill.
- Roth, P. L., Purvis, K. L., & Bobko, P. (2012). A metaanalysis of gender group differences for measures of job performance in field studies. *Journal of Management*, 38, 719–739.
- Rothstein, H. R., & Hopewell, S. (2009). Grey literature. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and metaanalysis* (2nd ed., pp. 103–125). New York: Russell Sage Foundation.
- Rotton, J., Foos, P. W., Van Meek, L., & Levitt, M. (1995). Publication practices and the file drawer problem: A survey of published authors. *Journal of Social Behavior and Personality*, 10, 1–13.
- Sanchez-Meca, J., Marin-Martinez, F., & Chacon-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in metaanalysis. *Psychological Methods*, 8(4), 448–467.
- Schmidt, F. L. (1992). What do data really mean? Research findings, metaanalysis, and cumulative knowledge in psychology. *American Psychologist*, 47, 1173–1181.
- Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1, 115–129.
- Schmidt, F. L., Le, H., & Oh, I. (2009). Correcting for the distorting effects of study artifacts in metaanalysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and metaanalysis* (2nd ed., pp. 317–333). New York: Russell Sage Foundation.
- Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, 470, 327.
- Shadish, W. R. (1996). Metaanalysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychological Methods*, 1, 47–65.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and*

quasi-experimental designs for generalized causal inference. New York: Houghton Mifflin.

Shea, B. J., Grimshaw, J. M., Wells, G. A. *et al.* (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology*, 7, 10–17.

Sheppard, B. H., Hartwick, J., & Warshaw, P. R. (1988). The theory of reasoned action: A metaanalysis of past research with recommendations for modifications and future research. *Journal of Consumer Research*, 15, 325–343.

Sidanius, J., Pratto, F., & Bobo, L. (1994). Social dominance orientation and the political psychology of gender: A case of invariance? *Journal of Personality and Social Psychology*, 67, 998–1011.

Snedecor, G. W., & Cochran, W. G. (1980). *Statistical methods* (7th ed.). Ames: Iowa State University Press.

Sommer, B. (1987). The file drawer effect and publication rates in menstrual cycle research. *Psychology of Women Quarterly*, 11, 233–242.

Sternberg, R. J. (1991). Editorial. *Psychological Bulletin*, 109, 3–4.

Stewart, L., Tierney, J., & Burdett, S. (2005). Do systematic reviews based on individual patient data offer a means of circumventing biases associated with trial publications? In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in metaanalysis: Prevention, assessment and adjustments* (pp. 261–286). New York: Wiley.

Stigler, S. M. (1986). *History of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and metaanalysis* (2nd ed., pp. 435–452). New York: Russell Sage Foundation.

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A metaanalysis. *Psychological Bulletin*, 121, 371–394.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.

Taras, V., Kirkman, B. L., & Steel, P. (2010). Examining the impact of culture's

- consequences: A three-decade, multilevel, meta-analytic review of Hofstede's cultural value dimensions. *Journal of Applied Psychology*, 95(3), 405–439.
- Thompson, K. N., & Schumacker, R. E. (1997). An evaluation of Rosenthal and Rubin's binomial effect size display. *Journal of Education and Behavioral Statistics*, 22(1), 109–117.
- Timm, N. H. (1975). *Multivariate analysis, with applications in education and psychology*. Belmont, CA: Brooks-Cole.
- Twenge, J. M. (2000). The age of anxiety? The birth cohort change in anxiety and neuroticism, 1952–1993. *Journal of Personality and Social Psychology*, 79(6), 1007–1021.
- Twenge, J. M., Gentile, B., DeWall, C. N., Ma, D., Lacefield, K., & Schurtz, D. R. (2010). Birth cohort increases in psychopathology among young Americans, 1938–2007: A cross-temporal metaanalysis of the MMPI. *Clinical Psychology Review*, 30, 145–154.
- Twenge, J. M., Konrath, S., Foster, J. D., Campbell, W. K., & Bushman, B. J. (2008). Egos inflating over time: A cross-temporal metaanalysis of the narcissistic personality inventory. *Journal of Personality*, 76, 875–901.
- Twenge, J. M., & Nolen-Hoeksema, S. (2002). Age, gender, race, socioeconomic status, and birth cohort difference on the children's depression inventory: A metaanalysis. *Journal of Abnormal Psychology*, 111(4), 578–588.
- United Nations Statistics Division (2012). International databases. Retrieved April 6 from <http://unstats.un.org/unsd/default.htm>.
- Valentine, J. C. (2009). Judging the quality of primary research. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and metaanalysis* (2nd ed., pp. 129–146). New York: Russell Sage Foundation.
- Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for metaanalysis. *Journal of Educational and Behavioral Statistics*, 35, 215–247.
- Viechtbauer, W. (2010). Conducting metaanalyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48.
- White, H. D. (2009). Scientific communication and literature retrieval. In H.

- Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and metaanalysis* (2nd ed., pp. 51–71). New York: Russell Sage Foundation.
- Wicker, A. W. (1969). Attitude versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, 25(4), 41–78.
- Wilson, D. B. (2009). Systematic coding. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and metaanalysis* (2nd ed., pp. 159–176). New York: Russell Sage Foundation.
- Winer, B. J., Brown, D. R., & Michels, K. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.
- Wood, W. (1987). Meta-analytic review of sex differences in group performance. *Psychological Bulletin*, 102, 53–71.
- Wood, W., & Eagly, A. H. (2009). Advantages of certainty and uncertainty. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and metaanalysis* (2nd ed., pp. 455–472). New York: Russell Sage Foundation.
- World Bank (2013). *Data*. Retrieved from <http://data.worldbank.org/>
- World Values Survey (2012). Retrieved April 6, 2012, from <http://www.worldvaluessurvey.org/>

* The preparation of this chapter was facilitated by U.S. Public Health Service grants K18 AI094581–01 and R01 MH58563–13 to Blair T. Johnson.

We thank Tania B. Huedo-Medina and Cleo Protogerou for their helpful comments on previous drafts of this chapter.

Correspondence should be directed to either Blair T. Johnson, Department of Psychology, University of Connecticut, Unit 1020, 406 Babbidge Road, Storrs, CT 06269–1020 (e-mail: blair.t.johnson@uconn.edu) or to Alice H. Eagly, Department of Psychology, Northwestern University, Swift Hall, 2029 Sheridan Road, Evanston, IL 60208–2710 (e-mail: eagly@nwu.edu).

¹ Strictly speaking, metaanalysis concerns only a statistical integration, the “analyses of analyses” that the term literally connotes. Nonetheless, in practice, reviews that include analyses of analyses are usually labeled metaanalyses, meaning more broadly the entire research synthesis process. For clarity, a *systematic review* is generally one that attempts to grade evidence relevant to a question; it may or may not include metaanalysis per se.

² Meta-analysts are wise to consider the potential coverage of the moderators planned for analyses (Card, 2012). Merely randomly sampling studies from the frame of available studies may leave some moderators relatively sparse at values of theoretical interest. In oversampling among extreme values on the moderator, stratified random samples maximize available moderator variance and thus make statistical tests more sensitive.

³ Similarly, use of an unstandardized outcome as the effect size (e.g., unstandardized regression slope or unstandardized mean difference) requires that each study assessed the phenomenon using the same operations. For example, Kirsch *et al.* (2008) used the unstandardized difference in improvement in depression scores as T because every study in their metaanalysis used exactly the same measure of depression.

⁴ In parallel, if the odds ratio is T , one might define values greater than one as positive and those smaller than one as negative.

⁵ A more subtle form of nonindependence occurs when samples within particular studies are related, such as husbands in one sample and wives in another, or when single investigators contribute more than one study. Current convention offers no satisfactory solution to this problem except to conduct sensitivity analyses to determine whether including dependent cases affects statistical inferences (Greenhouse & Iyengar, 2009) or to conduct individual participant metaanalyses that can directly accommodate the dependencies (Stewart, Tierney, & Burdett, 2005).

⁶ Because the corrections information may sometimes be correlated with moderator dimensions, it seems that the most defensible strategy is to use the corrections as moderators themselves so that model testing can incorporate both

types of information simultaneously and thus determine which aspects uniquely explain variation in the effect sizes.

⁷ Q and I^2 may also be defined using random-effects assumptions.

⁸ This PsycINFO search was performed on April 2, 2012, with “metaanalysis” in title, abstract, or keywords; AND Content Classification Code = social psychology, personality psychology, personality scales and inventories, political processes and political issues, or sex roles and women's issues; AND Document type = journal article, chapter, or dissertation.

Appendix A: Estimating Effect Sizes in Individual Studies

A comprehensive treatment of the formulas to convert primary-level statistics to effect sizes is beyond the scope of this chapter (see Card, 2012; Glass et al., 1981; Johnson, 1993; Lipsey & Wilson, 2001; Rosenthal, 1991). Here we offer only the most common transforms for deriving g , the standardized mean difference effect size. For producing r from various statistical reports, Glass *et al.* (1981) provided several useful formulas; alternatively, the standardized mean difference, g (see Table 26.2), may be calculated and transformed to r by this equation:

$$r = \frac{g}{\sqrt{g^2 + 4}}. \quad (26.1A)$$

Effect Sizes from Means and Standard Deviations

Table 26.2, line 1, shows the equation to transform two means and a standard deviation into an effect size, $(M_A - M_B) / SD_{pooled}$. Yet, there are many possible forms of the standard deviation that can appear in the dominator of the formula. To derive g from means and standard deviations in a between-subjects design, it is conventional to use the pooled standard deviation, SD ,

$$SD_{pooled} = \sqrt{\frac{(n_A - 1)(SD_A)^2 + (n_B - 1)(SD_B)^2}{n_A + n_B - 2}},$$

(26.2A)

where n_A and n_B are the number of observations in the two groups being compared, and SD_A and SD_B are their standard deviations (Glass et al., 1981). Thus, SD represents the square root of a “pooling” of the variances of the two groups and is an identical variability estimate to that obtained when an F - or t -test evaluates the difference between the means of the two groups.

For within-subjects designs, Becker (1988) recommended using the pretest SD as the denominator when pretest and posttest scores are compared. Other within-subjects comparisons may be calculated as between-subjects when cell standard deviations are available. Alternatively, SD_{pooled} can be replaced with SD_d , the standard deviation of the differences between paired observations,

$$SD_d = \sqrt{SD_A^2 + SD_B^2 - 2r_{EC} SD_A SD_B}, \quad (26.3A)$$

where r_{EC} is the correlation between the paired observations (e.g., Dunlap, Cortina, Vaslow, & Burke, 1996). This form of the SD is equivalent to the $\sqrt{MS_{Error}}$ term in a repeated measures analysis of variance or in a t -test, which will generally provide relatively liberal estimates of effect size. Most often all of the components of this formula are not provided, and a paired-observation t -test or a within-subjects F is given instead. As we indicate in the next subsection, these statistics may be directly converted into the effect size that has the standard deviation of the differences in its dominator.

As a rule, whenever possible, SD should be estimated only from the portion of each study's data entering into the effect size. For example, if the $M_A - M_B$ difference needs to be calculated within a level of another variable, SD should be estimated from the standard deviations given for participants within this level, if this information is available. Often, however, SD is available only pooled across all of the conditions of an experiment. If the SD pooled within the cells of the design is not available, but the report contains a standard deviation for the overall sample, it should be converted to the pooled SD by removing the variance resulting from the difference between M_A and M_B (e.g., Hedges & Becker, 1986; Johnson, 1993).

Effect Sizes from t - and F -values

Calculations of g can also be based on summary statistics. In the case of the t -test for independent groups,

$$t = \frac{M_A - M_B}{\sqrt{\frac{SD_A^2}{n_A} + \frac{SD_B^2}{n_B}}}. \quad (26.4A)$$

Rearrangement of the terms of this equation produces the following formula for calculating g :

$$g = t \sqrt{\frac{n_A + n_B}{n_A n_B}} \quad (26.5A)$$

Or, if $n_A = n_B$,

$$g = t \sqrt{\frac{2}{n}} = \frac{2t}{\sqrt{2n}}. \quad (26.6A)$$

Because $t = \sqrt{F}$ for a comparison of two groups, when the F results from a between-subjects design with unequal n ,

$$g = \sqrt{F \frac{n_A + n_B}{n_A n_B}}. \quad (26.7A)$$

Or, if $n_A = n_B$,

$$g = \sqrt{F \frac{2}{n}}, \quad (26.8A)$$

where n is the within-cell n (not the total N). If a within-subjects t (i.e., for paired observations) is reported,

$$(26.9A)$$

$$g = \frac{t}{\sqrt{n}}.$$

When a study reports an F for a two-groups within-subjects comparison,

$$g = \sqrt{\frac{F}{n}}. \quad (26.10A)$$

Note that because equations 26.9A and 26.10A assume a repeated measures error variance (see equation 26.2A), they generally will provide relatively large estimates of effect size.

F -values that derive from designs with three or more conditions require some special consideration. F -values that have more than one degree of freedom in the numerator cannot be directly converted into effect sizes because they do not directly gauge differences between individual means. Rather, a significant omnibus F -value implies that somewhere among the relevant means, one or more significant differences exist (see Judd, Yzerbyt, & Muller, Chapter 25 in this volume). Thus, for example, a significant F -value from a design that uses low, medium, and high levels of the independent variable must be decomposed in order to permit effect size derivations. If a linear contrast is reported, it will be equivalent to a comparison between the high and low levels. One could compare the means only for the high and low levels or also compare the medium level with the low and the high levels (e.g., Rhodes & Wood, 1992). Or, if the relation between the independent and dependent variables is expected to be linear, one could compute an F for the linear trend in the means and transform it into g (see Glass et al., 1981; Rosenthal & Rosnow, 1985). Of course, analysts should use the means in a particular study that would produce the most similar comparison to that used to represent the other studies in the sample. Treating studies' results in substantially different ways would introduce noise into the effect sizes in the database.

Similar issues arise in designs with two or more factors. In such instances, to make effect size comparisons more similar across the studies in a meta-analytic sample, some methodologists have recommended producing one-way designs by returning the effects of irrelevant factors to the error term of the ANOVA (Glass et al., 1981; Hedges & Becker, 1986; Morris & DeShon, 1997). This procedure should be seriously considered for individual-difference variables that were

crossed with the crucial independent variable in only some of the studies, because this source of variability would not have been removed from the error term in studies that did not assess these individual differences. When these irrelevant variables were instead manipulated, the decision is less straightforward, to the extent that researchers have created extreme conditions atypical of natural settings by means of powerful experimental manipulations. Variability stemming from extreme or atypical conditions would not be in the error term of typical studies. Therefore, adding sums of squares for such manipulated variables to the sum of squares error could greatly inflate these error terms in at least some instances and thus decrease the absolute magnitude of effect sizes based on these error terms. As Morris and DeShon (2002) concluded, in deciding whether to return irrelevant factors to the error term, analysts should keep as their goal the production of error terms that are based on the same sources of variability across the studies in the sample.

TABLE 26.1A. *Hypothetical analysis of variance summary tables (a) before reconstitution and (b) after returning factor B's sums of squares to the error term degrees*

Source	Sum of squares	Degrees of freedom	Mean squared error	<i>F</i>
(a) Before reconstituting				
A	430.33	1	430.33	15.22
B	200.12	1	200.12	7.08
A × B	43.55	1	43.55	1.54
Error	1,244.29	44	28.28	
(b) After reconstituting				
A	430.33	1	430.33	13.30
Error	1,487.96	46	32.35	

To illustrate how to return irrelevant factors to the error term, [Table 26.1A](#) contains a hypothetical ANOVA for a two-factor design. The top panel contains the ANOVA summary for the two factors. Suppose that Factor A is the focal independent variable, and that Factor B is a meta-analytically irrelevant variable. To represent the impact of Factor A on the dependent variable, the variation due to Factor B can be returned to the error term. This operation is performed by (a) adding the sum-of-squares due to Factor B and its interaction with Factor A to the error sum-of-squares and (b) adding the degrees of freedom due to Factor B and its interaction to the degrees of freedom for error. Once the sum-of-squares

for error has been divided by its new degrees of freedom, the square root of the resulting mean-square for error would be interpretable as the standard deviation pooled within the two levels of A , or $SD = \sqrt{MS_e}$. The result of this reconstitution of the error term appears in Panel b. In this example, g may be derived by converting the F -value that resulted from the reconstitution procedure, or it may be derived by dividing the difference between the means of Factor A by SD . Morris and DeShon (1997) presented other equations and examples of this strategy; Nouri and Greenberg (1995) presented techniques for use with more complex ANOVA designs (e.g., those that mix between-and within-subjects factors).

If the effects of the focal independent variable on the dependent variable are expected to change within the levels of another independent variable, separate effect sizes can be calculated within levels of the second independent variable, as we already mentioned above (see subsection “Multiple Reports from Individual Studies”). Specifically, as an alternative to representing the effect of the focal independent variable aggregated over this other variable (i.e., as a main effect), the analyst can partition each study on this other variable and represent the effect of interest within levels of this variable (i.e., as a simple main effect). When interactions are expected, simple main effects are the desired comparison, and the other, interacting variable can function as a moderator of the relation between the focal variables. As an example, Table 26.2A displays a 2×3 factorial design in which the focal independent variable (IV_{focal}) and a moderator variable ($IV_{\text{moderator}}$) serve as the factors. Suppose that we expect the effect of IV_{focal} on the dependent variable to change depending on the level of $IV_{\text{moderator}}$. To represent these contrasting expectations, a separate effect size must be derived for each level of $IV_{\text{moderator}}$. Thus, the first g would result from a comparison of the means from cells a and b , the second from cells c and d , and the third from cells e and f . To perform this calculation, it is necessary to obtain all cell means and either (a) the within-cell standard deviations, (b) the standard deviations for each relevant level of $IV_{\text{moderator}}$ (and transformed to SD_{pooled}), or (c) MS_e for the ANOVA. The MS_e can be recovered when all cells means are reported and at least one F -value is known for the dependent variable, even when the available F is not the most relevant to the analysts’ focal comparison (Johnson, 1993; Morris & DeShon, 1997). These calculations are facilitated if the source report contains a complete ANOVA table, but the components of the table can be estimated if the means, cell sizes, and one or more F -values are known (Johnson, 1993). Then, $SD = \sqrt{MS_e}$. Once this value or the standard

deviations are known, effect-size derivations continue as though each condition were a separate experiment.

Table 26.2A. A Hypothetical factorial design in which a focal independent variable is crossed with a moderator-independent variable

		IV _{Focal}	
		Level 1	Level 2
IV _{Moderator}	Level 1	Cell <i>a</i>	Cell <i>b</i>
	Level 2	Cell <i>c</i>	Cell <i>d</i>
	Level 3	Cell <i>e</i>	Cell <i>f</i>

Finally, *F*-values derived from multivariate analysis of variance (MANOVA), in which one or more independent variables were examined for their simultaneous influence on two or more dependent measures, should not be transformed into effect sizes if the dependent variable of interest was combined with other, irrelevant dependent variables (see Morrison, 1976; Timm, 1975). If several measures of the same conceptual dependent variable were combined in a multivariate analysis, however, the analyst might derive an effect size by taking the square root of the proportion of variance that the independent variable accounts for in the best linear combination of the dependent variables and treating this value as an *r* (see Tabachnick & Fidell, 1996, pp. 388–391, discussion of Wilk's Lambda), even if univariate *F*-values from ANOVAs are not available. However, because such effect sizes would be dependent on the exact set of dependent variables included in the multivariate analysis, some meta-analysts recommend against such procedures (Hunter & Schmidt, 1990).

This discussion of *t*- and *F*-values shows that complex statistical considerations can arise in translating source reports into effect sizes. Because of these potential complexities, a reviewer should never proceed to calculate effect sizes from an ANOVA without thoroughly understanding the design used for the data analysis. The reviewer would be well advised to diagram the design with the relevant *ns*. Because multiple error terms are common in the designs used in

experimental social psychology, it is easy to use the wrong error term for calculating the effect size. To prevent such errors, advanced ANOVA texts are invaluable (e.g., Myers & Well, 1991; Winer, Brown, & Michels, 1991). For reference purposes, meta-analysts may find it convenient to produce a packet of the clearest textbook descriptions of designs that occur often in their literatures.

Effect Sizes from r -values

Although r can be readily transformed to g ,

$$g = \frac{2r}{\sqrt{1-r^2}}, \quad (26.11A)$$

correlational reports often appear in a form other than r (see Carroll, 1961; Cohen & Cohen, 1983; Glass et al., 1981; Rosenthal, 1991, 1994). When r -values other than the product-moment variety are reported (e.g., biserial r , phi coefficient), they can usually be interpreted as product-moment rs , except when they are point-biserial rs . In this case, the meta-analyst would convert the point-biserial r into the biserial r , which approximates the product-moment r . If $n_A = n_B$ or when n_A is approximately n_B , $r_b = 1.253r_{pb}$, or, if $n_A \neq n_B$,

$$r_b = \frac{r_{pb}\sqrt{n_A n_B}}{\mu N}, \quad (26.12A)$$

where N is the total sample size, and μ is the ordinate of the unit normal distribution (i.e., the height of normal curve with surface equal to 1.0 at the point of division between segments containing n_A and n_B cases). Similarly, if a study reports t calculated based on any r -value, the t can be converted to a product-moment correlation using

$$r_b = \frac{r_{pb}\sqrt{n_A n_B}}{\mu N}. \quad (26.13A)$$

Whereas standardized regression weights (β) deriving from simple linear regressions are r -values and can be so interpreted, β s deriving from regressions with more than one predictor *cannot* be directly interpreted as r -values. The β -

value for a given predictor in a multiple regression equation is *adjusted* for the other independent variables present in the equation. In the case of suppressor variables (Cohen & Cohen, 1983), these adjustments can affect not only the value of β but also its sign, which could be reversed from the sign of the correlation between the two variables. Yet another problem with converting β -values to effect sizes is that under some circumstances β -values from multiple regression equations exceed $|1|$, whereas r -values never exceed $|1|$. For example, if Equation 26.11A is used with a β of 1.1, the denominator of the equation will be the square root of a negative number, -0.21 , which is an irrational mathematical operation. Therefore, as a general rule, in metaanalyses for which multiple regression results are the exception and other studies in the sample report statistics unadjusted for the other variables in the equation, multiple regression results should not be converted to effect sizes (see Hunter & Schmidt, 1990). Of course, before discarding a study because its findings were reported in a multiple regression, one should see whether a correlation matrix or comparable statistics appear in the report or could be obtained from its authors.

If many of the studies in a literature contain multiple regression equations that use the same conceptual independent variables to predict the same conceptual dependent variable, syntheses could pursue two strategies. One alternative is to examine how much variance (estimated by multiple R^2) was explained in the criterion variable by the set of predictor variables. For example, an analyst might examine each study to determine how much variance in intentions to perform a behavior was explained by the simultaneous impact of attitudes toward performing the behavior and normative expectations about the behavior (see Sheppard, Hartwick, & Warshaw, 1988). Hedges and Olkin (1985, p. 239) provide an alternative strategy that relies directly on the β s and their sample sizes to produce an aggregate weighted beta-weight.

Effect Sizes from Chi-square Values

Chi-square (χ^2) values are sometimes used to test for the frequency with which groups meet some criterion or to test for the association between two variables (Hays, 1988). If the χ^2 results from a 2×2 classification table linking a predictor (X) to the outcome (Y), then r can be calculated:

$$r_{\phi} = \sqrt{\frac{\chi^2}{n}}, \quad (26.14A)$$

where r_ϕ is a phi coefficient and approximates the product-moment r and can be converted to g :

$$g = \frac{2r}{\sqrt{1-r^2}}. \quad (26.15A)$$

Note: that if there is more than 1 degree of freedom in the χ^2 value, it cannot be directly converted into an effect size because the χ^2 may describe a nonlinear pattern. It may be possible to compute χ^2 for an appropriate 2×2 table based on the proportions of the relevant groups that meet a criterion (see the next subsection). If the data for these recomputations are not available, the study result cannot be used to derive an effect size.

Effect Sizes from Proportions Meeting a Criterion

In some designs, the proportion of individuals in one group (p_E) who meet a given criterion is compared with the proportion of individuals in another group (p_C) who meet it. For example, the proportion of people who help another person in one experimental condition can be compared to the proportion of people who help in another condition (see Eagly & Crowley, 1986). Although these proportions can be transformed into an effect size by using a probit transformation (Glass et al., 1981) or by treating the proportions as means (Snedecor & Cochran, 1980), the most efficient solution is to use the Cox transformation of the odds ratio gauging the effect size (see [Table 26.2A](#), line 8),

$$g_{\text{Cox}} = \frac{\text{LOR}}{1.65}, \quad (26.16A)$$

where LOR is the logged odds ratio (Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003). Note that this equation assumes that the proportions are in relation to the study's unit of analysis, which usually is the numbers of persons. The equations do not apply to proportions that represent values of dependent variables assessed for each unit of analysis. For example, if each participant's helping were assessed by a self-report of the proportion of occasions on which he or she helped, these data would produce an effect size by equations that use the variability of these proportions (e.g., [Table 26.2](#), line 1) rather than [Equation 26.16A](#).

Effect Sizes from Probabilities Associated with Inferential Statistics

Source reports sometimes contain only a p -value associated with the critical effect (e.g., $p = .0439$), which can be used to calculate an effect size if the direction of the finding and the sample size (n) are known. To do so, the analyst would use a statistical package's (e.g., SAS, IMSL, SPSS, Stata) or a spreadsheet's inverse probability distribution functions, which provide an exact solution of a test statistic from p . For example, SAS provides BETAINV, which yields F from p and df , after which the F can be converted to g using Equations 26.7A through 26.10A (assuming that the F compares the means of only two groups). Obviously, an exact p allows an excellent estimate of a test statistic and therefore of g . Conversely, a level p (e.g., $p < .05$) gives a poorer estimate, because it would ordinarily be treated as exactly the p level given (e.g., $p < .01$ would be understood as $p = .01$). The mere statement that a finding is “significant” can be treated as $p = .05$ in studies that apparently use the conventional $p < .05$ rule for determining significance and indicate the direction of the effect, but the effect sizes estimated on this basis may be quite inaccurate (Ray & Shadish, 1996). Finally, reports often differ in whether a one-tailed or two-tailed probability level is reported; if no information is provided, the convention is that the study authors have used a two-tailed test.

Author Index

Aarts, H. [324](#)
Abakoumkin, G. [211](#)
Abarbanel, H. D. I. [271](#), [272](#)
Abboud, H. [205](#)
Abelson, R. P. [27](#), [30](#), [33](#), [34](#), [36](#), [40](#), [41](#), [44](#), [45](#), [696](#)
Abraham, F. D. [258](#)
Abraham, R. [258](#)
Abrams, R. L. [322](#)
Abramson, A. [222](#), [227](#)
Abramson, P. [431](#)
Acharya, T. [270](#)
Achee, J. [254](#), [262](#), [265](#)
Ackerman, R. A. [589](#)
Adams, C. M. [134](#)
Adams, G. [84](#), [85](#), [86](#)
Aderman, D. [23](#)
Adisetiyo, V. [132](#)
Adjali, I. [294](#)
Adler, M. [210](#)
Adolf, D. [151](#)
Adorno, T. W. [424](#)
Affleck, G. [373](#), [377](#), [384](#)
Agosta, F. [147](#)
Agosta, S. [299](#)
Ahern, G. L. [241](#)
Ahlawat, K. S. [539](#)
Ahlgren, A. [413](#)
Aiken, L. S. [56](#), [57](#), [59](#), [60](#), [62](#), [64](#), [521](#), [522](#), [523](#), [665](#), [667](#), [668](#), [669](#), [690](#)
Ajzen, I. [298](#), [481](#)
Akhtar, O. [296](#), [412](#)
Akinola, M. [224](#), [242](#)
Albarracín, D. [295](#), [694](#)
Albert, M. [132](#)

Alcañiz, M. [225](#), [231](#)
Alderman, H. [409](#)
Alexander, R. A. [209](#), [423](#)
Algoe, S. B. [445](#)
Allen, C. [255](#)
Allen, J. J. B. [221](#), [237](#), [241](#)
Allen, K. [446](#)
Allen, M. T. [112](#), [113](#), [115](#)
Allen, R. [221](#)
Allen, T. A. [303](#)
Allen, T. J. [301](#)
Alliger, G. M. [383](#)
Allison, D. B. [65](#), [176](#)
Allison, P. D. [363](#), [632–633](#)
Allport, F. H. [211](#), [313](#)
Allport, G. W. [114](#)
Almeida, D. M. [383](#), [385](#), [387](#)
Almli, C. R. [132](#)
Alpert, H. [161](#)
Alquist, J. L. [2](#)
Altermatt, T. W. [195](#), [199](#)
Altman, I. [191](#), [192](#), [211](#), [350](#)
Alvaro, E. M. [11](#)
Alwin, D. F. [408](#), [423](#), [424](#), [425](#)
Amazeen, P. G. [262](#)
Ambady, N. [254](#), [267](#), [268](#)
Amelsvoort, T. V. [226](#)
American National Election Studies [684](#)
American Psychological Association [476](#), [683](#), [697](#)
Amichai-Hamburger, Y. [445](#)
Amodio, D. M. [142](#), [147](#)
Amoss, R. T. [228](#)
Andersen, B. L. [513](#)
Anderson, A. B. [198](#)
Anderson, B. [431](#)
Anderson, B. F. [197](#)
Anderson, C. A. [589](#)
Anderson, E. [226](#), [227](#), [239](#)
Anderson, J. C. [526](#)

Anderson, J. R. [327](#), [331](#)
Anderson, M. L. [256](#)
Andersson, J. [147](#)
Andrews, B. [378](#)
Andrews-Hanna, J. R. [231](#)
Andrich, D. [558](#)
Angleitner, A. [491](#)
Angrist, J. D. [55](#), [56](#), [88](#)
Angst, J. [678](#)
Angus, D. [275](#)
Anscombe, F. J. [616](#), [618](#)
Ansfield, M. E. [336](#)
Anthony, B. J. [119](#)
Appenzeller, T. [254](#)
Arch, J. [135](#)
Arcuri, L. [295–296](#)
Argote, L. [201](#)
Ariely, D. [445](#), [447](#)
Arkin, R. [6](#), [350](#)
Armeli, S. [377](#), [384](#), [453](#)
Arminger, G. [547](#)
Aron, A. [84](#), [348](#), [457](#)
Aron, A. R. [124](#), [147](#)
Aron, E. N. [84](#), [348](#), [457](#)
Aronoff, J. [194](#)
Aronson, E. [2](#), [17](#), [21](#), [31](#), [44](#), [60](#), [83](#), [474–475](#), [489](#)
Aronson, J. [110](#)
Aronson, J. A. [146](#)
Arrow, H. [201](#), [254](#), [255](#), [256](#)
Arunachalam, V. [210](#)
Arvey, R. [41](#)
Ary, D. [53](#)
Asch, S. E. [11](#), [85](#), [190](#), [253](#), [375](#)
Ascoli, G. A. [144](#)
Asendorpf, J. B. [286](#)
Ashburner, J. [123](#), [130](#)
Asparouhov, T. [573](#), [586](#), [587](#)
Assenheimer, J. S. [537](#)
Atienza, A. A. [381](#)

Atkin, R [190](#), [192](#)
Atkins, D. C. [379](#), [383](#)
Atkinson, A. C. [612](#), [613](#)
Atkinson, R. C. [516](#), [518](#)
Avendano, M. [447](#)
Axelrod, R. [267](#)
Axinn, W. G. [182](#)

Baas, M. [680](#)
Babb, C. [151](#)
Babbie, E. R. [406](#)
Babcock, J. [445](#)
Babson, K. A. [223](#)
Baccaro, L. [599](#)
Bachman, J. G. [183](#)
Bachorowski, J. A. [227](#), [234](#), [237](#), [238](#)
Back, K. [189](#), [191](#)
Back, M. D. [599–600](#), [602](#)
Baecke, S. [151](#)
Bagozzi, R. P. [445](#)
Baier, C. J. [161](#)
Bailenson, J. [225](#), [231](#)
Bailenson, J. N. [231](#)
Bakeman, R. [204](#), [270](#), [351](#), [354](#), [355](#), [356–357](#), [359](#), [360](#), [362](#), [363–364](#)
Baker, C. I. [144](#)
Baker, C. M. [135](#)
Baker, R. P. [415](#)
Baker, S. G. [54](#)
Balasubramaniam, R. [273](#)
Baldueza, J. [423](#)
Baldwin, S. A. [603–604](#), [605](#)
Bales, R. F. [189](#), [198](#)
Balliet, D. [683](#)
Ballinger, G. A. [362](#)
Balluerka, N. [552](#)
Balodis, I. M. [128](#)
Balota, D. A. [333](#)
Banaji, M. R. [124](#), [125](#), [285](#), [314](#), [334](#), [445](#), [452](#), [464](#), [465](#), [466](#)
Bandalos, D. L. [633–634](#)

Bandettini, P. A. [126](#), [147](#)
Bandilla, W. [461](#)
Bandura, A. [190](#), [337](#)
Banich, M. T. [124](#), [148](#)
Banks, W. [348](#)
Banse, R. [222](#), [289](#)
Bar-Anan, Y. [289](#)
Barch, D. M. [151](#)
Barchard, K. A. [466](#)
Barcikowski, R. S. [57](#)
Bard, E. G. [269](#)
Bargh, J. A. [4](#), [34](#), [146](#), [223](#), [228](#), [235](#), [284](#), [311](#), [312](#), [314](#), [315](#), [317](#), [318](#), [319–322](#), [323–324](#), [326](#), [327](#), [329](#), [330](#), [332](#), [333](#), [335](#), [337](#), [343](#), [344](#)
Barndollar, K. [323](#)
Barner-Rasmussen, W. [423](#)
Barnes-Holmes, D. [291–292](#)
Barnes-Holmes, Y. [291–292](#)
Barnow, L. S. [68](#)
Baron, R. M. [14](#), [44](#), [60](#), [255](#), [262](#), [407](#), [519](#), [520](#), [523](#), [580](#), [654](#), [656](#), [658](#), [662](#), [664](#)
Barouei, J. [236](#)
Barrett, L. F. [220](#), [221](#), [222](#), [223](#), [225](#), [226](#), [227](#), [228](#), [230](#), [231](#), [232](#), [235](#), [236](#), [237](#), [238](#), [240](#), [241](#), [242](#), [243](#), [378](#), [379](#), [577](#)
Barrick, M. R. [209](#)
Barry, R. A. [446](#)
Barsalou, L. W. [223](#), [228](#), [229](#), [238](#)
Barta, W. D. [390](#)
Bartholomew, K. [200](#)
Bartlett, M. Y. [224](#), [230](#), [235](#)
Barton, S. [258](#)
Bartoshuk, L. M. [243](#)
Baruch, R. F. [28](#)
Barzantny, C. [423](#)
Bass, R. T. [431](#)
Bassingthwaighe, J. B. [275](#)
Bates, D. [586](#)
Bator, R. [84](#)
Bator, R. J. [325](#)
Battaglia, M. P. [414](#)

Baucom, B. R. [367](#)
Baucom, D. H. [351](#)
Bauer, D. J. [393–394](#), [525](#), [552](#), [668](#)
Bauer, M. S. [254](#)
Baumann, J. [230](#), [235](#)
Baumeister, R. F. [2](#), [124](#), [137](#), [146](#), [384](#), [447](#)
Baxter, P. M. [683](#)
Beach, R. [409](#)
Beale, E. M. L. [627](#)
Beall, A. C. [231](#)
Bearden, W. Q. [424](#)
Beatty, P. C. [428](#)
Bechtoldt, H. P. [537](#)
Beck, A. T. [242](#), [312](#)
Becker, B. J. [685](#), [686](#), [690](#), [694](#)
Becker, P. [499](#)
Beckett, Sean [409](#)
Beckmann, C. F. [131](#)
Beckmann, N. [376](#)
Beek, P. J. [262](#)
Beem, A. L. [178](#)
Beer, J. S. [123](#), [128](#), [149](#)
Begg, C. B. [693](#)
Begin, J. [364](#)
Behrend, T. S. [463](#)
Behrens, T. E. J. [131](#)
Behrman, J. R. [409](#)
Beimer, P. P. [431](#)
Belden, A. [151](#)
Bellgowan, P. S. F. [144](#)
Belmaker, R. H. [176](#)
Beltz, B. C. [256](#)
Bem, D. J. [30](#), [190](#), [427](#)
Bemston, G. G. [123](#)
Bench, S. W. [228](#), [229](#)
Benet, V. [420](#)
Benet-Martínez, V. [493](#), [498–499](#)
Benjamin, B. [178](#)
Benjamin, J. [176](#)

Benjamin, Y. [136](#)
Bennett, C. M. [147](#), [148](#)
Bennett, E. R. [176](#)
Bennett, M. E. [384](#)
Bennett, S. M. [234](#)
Benson, K. M. [597](#)
Benson, P. R. [364](#)
Bente, G. [127](#)
Bentler, P. M. [68](#), [479](#), [491](#), [513](#), [541](#)
Berdahl, J. L. [256](#)
Berenson, J. [23](#)
Berenson, R. [23](#)
Berent, M. K. [424](#), [425](#)
Berg, C. A. [534](#)
Berglas, S. [110](#)
Bergman, C. [603–604](#)
Berinsky, A. J. [91](#), [425](#)
Berk, R. A. [69](#)
Berkman, E. T. [134](#), [135](#), [144](#), [146](#), [149](#), [150](#), [151](#), [376](#), [384](#)
Berkowitz, L. [21](#), [50](#), [589](#)
Berman, H. J. [84](#), [377](#)
Berman, J. J. [64](#)
Berman, W. H. [604](#)
Bernard, H. R. [381](#)
Berntson, G. G. [102](#), [104](#), [106](#), [139](#), [224](#), [230](#), [232](#), [233](#), [235](#), [239](#), [240](#), [269](#)
Berridge, K. C. [235](#)
Berry, A. [375](#), [379](#), [383](#)
Berry, C. M. [689](#)
Berry, J. W. [498](#)
Berscheid, E. [189](#), [380](#)
Bertrand, D. [447](#)
Bertrand, S. [447](#)
Bertuglia, C. S. [259](#)
Beruvides, M. G. [210](#)
Berwick, J. [128](#)
Best, N. [560](#)
Bethlehem, J. [415](#)
Bettencourt, B. A. [684](#), [696](#)
Betts, K. R. [200](#)

Betz, A. L. [381](#)
Beurpré, M. G. [222](#)
Bickman, L. [74](#)
Biemer, P. [428](#)
Biesanz, J. C. [602](#)
Biglan, A. [53](#)
Bilder, R. M. [148](#)
Billiet, J. [431](#)
Billing, M. [191](#)
Binning, K. R. [445](#)
Birbaumer, N. [230](#)
Birn, R. M. [126](#)
Birnbaum, G. E. [383](#), [384](#)
Birnbaum, M. H. [444](#), [458](#), [459](#), [466](#)
Bischoping, K. [427](#)
Bishop, G. F. [430](#)
Biswal, B. B. [134–135](#)
Bjorkner, E. [227](#), [237](#)
Bjorner, J. B. [563](#)
Black, C. [53](#)
Black, M. [367](#)
Blackwell, L. [88](#)
Blaine, D. [176](#)
Blair, G. [89](#)
Blair, I. [334](#)
Blair, J. [428](#)
Blalock, H. M. [407](#), [408](#)
Blanchard, C. D. [238](#)
Blanchard, C. M. [423](#)
Blanchard, R. J. [238](#)
Blaney, P. H. [381](#)
Blank, A. [149](#)
Blanton, H. [300](#)
Blascovich, J. [102](#), [103](#), [104](#), [105](#), [107](#), [110](#), [111](#), [112](#), [113](#), [114](#), [115](#), [118](#), [119](#),
[120](#), [225](#), [229](#), [231](#), [240](#), [350](#)
Blasi, A. [485](#)
Bliss, C. A. [270](#)
Bliss-Moreau, E. [147](#), [220](#), [222](#), [227](#), [228](#), [232](#), [241](#), [242](#), [243](#)
Blitstein, J. L. [57](#)

Block, J. H. [482](#), [490](#), [493](#)
Bloxom, B. M. [237](#)
Bluemke, M. [300–301](#)
Blumberg, S. J. [419](#)
Bobko, P. [688](#)
Bobo, L. [408](#), [678](#)
Boccaro, N. [255](#)
Bock, R. D. [539](#), [559](#), [560](#)
Bodenhausen, G. V. [294](#), [327](#), [490](#)
Bodner, T. E. [627](#)
Bodurka, J. [147](#)
Boehnke, M. [177](#)
Boer, D. [684](#)
Bogaert, A. F. [420](#)
Bogart, K. R. [445](#)
Bogart, L. M. [208](#), [380](#), [384](#)
Boh, L. [207](#)
Bohrer, R. E. [272](#)
Boker, S. M. [272](#), [586](#)
Boldry, J. G. [378](#), [603](#)
Boles, S. [291–292](#)
Bolger, N. [86](#), [373](#), [376](#), [379](#), [380](#), [383](#), [387](#), [390](#), [393](#), [394](#), [512](#), [523](#), [525](#), [581](#),
[605](#), [656](#), [658](#)
Bolger, K. A. [230](#)
Bolhuis, J. E. [203](#)
Bolker, B. [586](#)
Bollen, K. [232](#)
Bollen, K. A. [480](#), [491](#), [512](#), [513](#), [520](#), [525](#), [526](#), [541](#), [658](#)
Boly, M. [151](#)
Bomhra, A. [176](#)
Bonabeau, E. [255](#)
Bond, C. F., Jr. [589](#), [605](#), [678](#), [690](#)
Bond, M. H. [498](#)
Bond, R. [682](#)
Bond, R. N. [34](#), [316](#), [320–321](#)
Boninger, D. S. [425](#)
Bonito, J. A. [598](#)
Bonser, I. M. [383](#)
Boomsma, D. I. [178](#)

Bootsmiller, B. [53](#)
Boquet, A. J. [113](#)
Bordin, E. S. [603](#)
Borenstein, M. [693](#), [694](#), [698](#)
Borg, I. [515–516](#), [517–518](#)
Borg, M. J. [431](#)
Borgatta, E. F. [201](#)
Boring, E. G. [312](#)
Bork, A. [130](#)
Borkowski, W. [259](#)
Borman, G. D. [693](#), [694](#)
Bormann, C. [327](#)
Bornstein, R. [322](#)
Borofsky, L. A. [135](#), [150](#)
Boruch, R. F. [50](#), [53](#), [54](#), [57](#)
Borzovsky, P. [53](#)
Boscan, P. [239](#)
Bosker, R. J. [57](#), [575](#), [586](#), [592](#)
Bosnell, R. [147](#)
Bosnjak, M. [461](#)
Botella, J. [692](#)
Botteron, K. [132](#)
Bouchard, S. [225](#)
Bouchard, T. J. [159](#), [161](#)
Boulton, A. [586](#)
Bousfield, A. K. [330](#)
Bousfield, W. A. [330](#)
Bowdle, B. F. [224](#), [230](#), [419](#)
Bowen, B. D. [406](#)
Bower, G. H. [330](#), [331](#)
Bowlby, J. [396](#)
Bowles, S. [201](#)
Bowman, F. D. B. [134](#)
Box, G. E. P. [60](#), [67](#), [334](#), [620](#)
Box-Steffensmeier, J. M. [410](#)
Boyce, A. [446](#)
Boyd, R. W. [102](#)
Boynton, M. H. [678](#), [682](#), [685](#), [690](#), [691](#)
Brackett, D. J. [225](#)

Bradburn, N. M. [422](#), [425](#), [426](#), [429](#), [431](#)
Bradbury, T. [347](#)
Bradley, M. M. [115](#), [118](#), [119](#), [220](#), [221](#), [222](#), [223](#), [226](#), [227](#), [228](#), [237](#), [239](#), [240](#)
Bradshaw, C. P. [49](#), [59](#)
Brammer, M. J. [147](#), [226](#)
Branch, L. G. [198](#)
Brandt, F. [128](#), [224](#), [230](#)
Brannon, L. A. [409](#)
Branscombe, N. R. [327](#), [445](#)
Brant-Zawadzki, M. [129](#)
Braucht, G. N. [53](#)
Brault, M. [364](#)
Braver, S. L. [86](#)
Bray, R. M. [192](#), [198](#)
Brechan, I. [295](#), [694](#)
Brehm, J. [408](#), [417](#), [418](#)
Brehm, J. W. [325](#)
Breivik, E. [526](#)
Brekke, N. [322](#)
Brennan, L. [242](#)
Brenner, M. [329](#)
Brestan-Knight, E. [348](#)
Brett, J. M. [592](#), [598](#), [654](#)
Brett, M. [136](#), [147](#)
Breuning, M. [408](#)
Brewer, M. B. [21](#), [22](#), [191](#), [204](#), [326](#), [381](#), [422](#)
Brewin, C. R. [378](#)
Brick, T. [586](#)
Brick, J. M. [418](#)
Brickman, P. [4](#)
Bridge, R. G. [409](#)
Briggs, N. E. [506](#), [507](#), [582](#), [585](#)
Briggs, S. [482](#), [489–490](#)
Brillinger, D. R. [609](#)
Briner, R. B. [381](#)
Brinthaup, T. M. [481](#)
Brislin, R. W. [498](#)
Brissette, I. [381](#)
Bristol, T. [208](#)

Broadbent, D. E. [315](#)
Brock, R. L. [446](#)
Broderick, J. E. [381](#), [387](#), [389](#)
Brody, N. [321](#)
Bromley, S. [87](#)
Brow, A. [88](#)
Brown, A. [422–423](#)
Brown, B. B. [210](#)
Brown, C. H. [366–367](#)
Brown, G. G. [147](#)
Brown, J. D. [336](#), [337](#), [489](#)
Brown, K. D. [432](#)
Brown, K. W. [254](#), [384](#)
Brown, R. [30](#)
Brown, R. L. [633–634](#)
Brown, S. C. [330](#)
Brown, T. A. [566](#), [576](#)
Browne, M. W. [170](#), [507](#), [508](#), [511](#), [513](#), [515](#), [541](#), [547](#)
Brubaker, T. H. [677](#)
Bruckman, A. [465](#), [466](#)
Brumbaugh, C. C. [375](#), [383](#)
Brunell, A. B. [589](#)
Bruner, J. S. [313](#), [317](#)
Brunet, E. [134](#)
Brunswik, E. [31](#), [33](#)
Bryant, K. J. [548](#)
Bryk, A. S. [57](#), [162](#), [366](#), [392](#), [524](#), [525](#), [575](#), [592](#), [598](#)
Buchanan, T. [446](#)
Buchels, C. [230](#)
Buchner, A. [53](#)
Buckley, K. E. [589](#)
Buckley, T. [211](#)
Buckner, R. L. [126](#), [231](#)
Bucy, E. P. [581](#)
Buder, E. H. [254](#), [259](#), [263](#)
Buechel, C. [135](#)
Bug, W. J. [144](#)
Buhrmester, M. [462](#), [463](#)
Bui-Wrzosinska, L. [262](#)

Buja, A. [508](#)
Bullmore, E. T. [226](#)
Bullock, J. G. [44](#), [425](#), [664](#)
Bundy, R. [191](#)
Bunge, S. A. [124](#), [146](#)
Bunz, H. [273](#)
Burdett, S. [687–689](#), [695–696](#)
Burger, K. S. [127](#)
Burisch, M. [482](#), [496](#)
Burke, B. L. [444](#)
Burke, M. J. [689](#)
Burklund, L. [135](#)
Burks, A. T. [432](#)
Burrow, A. L. [379](#), [380](#), [383](#)
Burrows, L. [318](#), [323](#), [335](#), [343](#)
Burt, S. A. [160](#)
Busa, E. [132](#)
Busemeyer, J. R. [34](#), [669](#)
Bush, A. L. [379](#), [384](#)
Bush, J. W. [424](#)
Bush, L. K. [40](#)
Bushery, J. [427](#)
Bushman, B. J. [688](#), [690](#), [695](#)
Busk, L. K. [233](#)
Butera, F. [655](#), [659](#)
Butler, E. A. [364](#), [366](#)
Butler, J. L. [329](#)
Butler, P. [66](#), [67](#)
Buttram, R. T. [327](#)
Buzsaki, G. [137](#)
Bylsma, W. H. [379](#)
Byrne, B. M. [481](#)
Byrne, D. [409](#)

Cabooter, E. [424](#)
Caceres, A. [147](#)
Cacioppo, J. T. [19](#), [21](#), [41](#), [102](#), [103](#), [104](#), [106](#), [115](#), [116](#), [119](#), [123](#), [138](#), [139](#),
[224](#), [226](#), [230](#), [232](#), [233](#), [234](#), [235](#), [239](#), [240](#), [243](#), [269](#), [405](#), [487–488](#), [520](#)
Cafri, G. [682](#)

Cai, L. [515](#), [561](#)
Cain, V. S. [380](#)
Calder, B. J. [65](#), [67](#)
Calhoun, V. D. [138](#)
Camazine, S. [255](#)
Camerer, C. F. [128](#)
Campbell, C. [605](#)
Campbell, D. T. [11](#), [12](#), [16](#), [18](#), [21](#), [22](#), [27](#), [28](#), [35](#), [36](#), [39](#), [40](#), [43](#), [45](#), [49](#), [50](#), [54](#),
[57](#), [59](#), [60](#), [61](#), [62](#), [64](#), [73](#), [74](#), [81](#), [83](#), [86](#), [87](#), [88](#), [89](#), [90](#), [93](#), [109](#), [188](#), [395](#),
[489](#), [490](#), [491](#), [543](#), [680](#)
Campbell, D. W. [145](#)
Campbell, J. D. [379](#)
Campbell, J. P. [360](#)
Campbell, L. [378](#)
Campbell, P. L. [236](#)
Campos, B. [347](#)
Canabal, A. [423](#)
Canavello, A. [379](#), [384](#)
Cannell, C. F. [200](#), [427](#), [430](#)
Cannon, W. B. [239](#)
Canto, A. Y. [210](#)
Cantor, D. [417](#)
Cantor, N. [323](#), [337](#)
Capideville, C. S. [225](#), [231](#)
Capitman, J. A. [317](#)
Caplan, R. D. [59](#)
Cappelleri, J. C. [64](#)
Caprariello, P. A. [445](#)
Card, N. A. [580](#), [678](#), [682](#), [684](#), [689](#), [698](#)
Carello, C. [255](#), [263](#), [273](#)
Carey, G. [161](#)
Carey, M. P. [681](#)
Caria, A. [230](#)
Caridakis, G. [270](#)
Carley, K. M. [201](#)
Carlin, J. B. [164](#)
Carlsmith, J. M. [31](#), [44](#)
Carlsmith, K. [60](#), [83](#)
Carlson, A. [151](#)

Carlson, M. [684](#)
Carlston, D. E. [331](#), [378](#)
Carmichael, C. L. [445](#)
Carnagey, N. L. [589](#)
Carney, M. A. [377](#), [384](#)
Carnot, C. G. [409](#), [425](#)
Caron, R. [269](#)
Carpenter, M. [324](#)
Carpentier, F. D. [54](#)
Carrington, P. J. [202](#)
Carroll, J. B. [534](#)
Carroll, J. D. [516](#), [518](#)
Carroll, J. E. [236](#)
Carroll, J. L. [62](#), [64](#)
Carroll, J. M. [243](#)
Carson, R. T. [425](#)
Carstensen, L. L. [375](#)
Carter, C. S. [142](#)
Carter, L. E. [223](#)
Carter, M. [204](#)
Carterette, E. C. [516](#), [518](#)
Carver, C. S. [323](#)
Case, T. I. [236](#)
Casey, B. J. [147](#)
Casey, M. A. [208](#)
Caspi, A. [380](#)
Castanier, C. [225](#)
Castel, A. D. [149](#)
Castellano, G. [270](#)
Castrop, F. [230](#)
Cattell, R. B. [160](#), [480](#), [490](#), [507](#), [534](#), [539](#), [547](#)
Cavallaro, L. A. [229](#)
Ceballos-Baumann, A. O. [230](#)
Cela-Conde, C. J. [255](#)
Chaiken, S. [316](#), [318](#), [332](#), [333](#), [481](#)
Chaires, W. M. [313](#)
Cham, H. [57](#)
Chambers, J. M. [610](#)
Chambers, W. [323](#)

Chance, J. E. [198](#)
Chance, Z. [445](#), [447](#)
Chandler, J. [463](#)
Chang, K.-M. [151](#)
Chang, L. [416](#), [421](#), [431](#)
Chanley, V. A. [408](#)
Chanowitz, B. [149](#)
Chaplin, W. F. [496](#)
Chapman, I. [422–423](#)
Chapman, K. E. [445](#), [453](#)
Charney, E. [180](#)
Chartrand, E. [350](#), [364](#)
Chartrand, T. L. [318](#), [321](#), [324](#), [329](#), [330](#), [343](#), [344](#), [350](#)
Chatfield, C. [67](#)
Chatham, C. H. [124](#), [148](#)
Chatzisarantis, N. L. D. [679](#)
Cheek, J. M. [487](#)
Cheever, N. A. [445](#)
Chemero, A. [256](#)
Chemers, M. M. [577](#)
Chemy, S. S. [178](#)
Chen, C.-T. [560](#)
Chen, J. [114](#), [323](#)
Chen, M. [235](#), [318](#), [323](#), [335](#), [343](#)
Chen, M. M. [424](#)
Chen, S. [316](#), [633–634](#)
Chen, W.-H. [561](#)
Cheng, K. [23](#)
Cheng, S. [430](#)
Chertkoff, J. M. [191](#)
Cheung, C. K. T. [146](#), [452](#)
Cheung, G. W. [547](#), [579](#), [581](#)
Cheung, M. N. [272](#)
Cheung, M. W. L. [659](#)
Chew, B. [379](#)
Chiaburu, D. S. [689](#)
Chlebus, P. [147](#)
Choi, D.-H. [413](#)
Choi, J. [201](#)

Choi, W. [146](#), [452](#)
Chopra, M. [90](#)
Chopra, S. [134](#)
Christensen, A. [349](#), [367](#), [394](#)
Christian, L. M. [422–423](#), [430](#)
Chu, W. W. [148](#)
Chuang, Y. C. [425](#)
Chumbley, J. R. [136](#)
Church, A. T. [498](#)
Chute, D. L. [205](#)
Cialdini, R. B. [81](#), [82](#), [83](#), [84](#), [86](#), [88](#), [325](#), [377](#), [430](#), [445](#)
Cicarelli, V. G. [68](#)
Ciccarelli, O. [147](#)
Cicchetti, D. [482](#)
Cicchetti, D. V. [356](#), [358](#), [359](#)
Cicchetti, P. [238](#)
Cillessen, A. H. N. [201](#), [598](#), [602](#)
Cingranelli-Richards Human Rights Project [684](#)
Clare Kelly, A. M. [134–135](#)
Clare, D. A. [424](#)
Clark, M. S. [334](#)
Clark, C. Z. F. [415](#)
Clark, F. [423](#)
Clark, J. K. [524](#)
Clark, L. A. [124](#), [237](#), [375](#), [377](#), [379](#), [383](#), [537](#), [635](#)
Clark, M. [146](#)
Clark, M. H. [73](#)
Clark, M. S. [381](#), [589](#)
Clark, T. G. [177](#)
Clark-Carter, D. [443](#)
Clarke, M. [449](#), [454](#), [455](#), [456](#), [457–458](#), [461](#), [464](#)
Clatts, M. [73](#)
Cleary, T. A. [547](#), [548](#)
Cleeremans, A. [318](#)
Cleveland, W. S. [63](#), [610](#), [615](#)
Cliff, N. F. [475](#)
Clinton, J. D. [409](#), [419](#)
Clogg, C. C. [541](#), [547](#)
Cloninger, C. R. [176](#)

Clore, G. L. [225](#), [243](#), [335](#), [411](#)
Coan, J. A. [127](#), [146](#), [221](#), [232](#)
Coberley, C. [53](#)
Coch, L. [190](#)
Cochran, W. G. [51](#)
Codispoti, M. [239](#), [240](#)
Cohen, D. [419](#)
Cohen, G. L. [89](#), [90](#)
Cohen, J. [30](#), [35](#), [37](#), [42](#), [51](#), [205](#), [356](#), [431](#), [443](#), [465](#), [466](#), [482](#), [499](#), [521](#), [522](#),
[523](#), [542](#), [658](#), [664–665](#), [668](#), [669](#), [670](#), [686](#), [690](#), [696](#)
Cohen, J. D. [136](#), [146](#), [147](#), [148](#)
Cohen, L. L. [379](#)
Cohen, M. D. [267](#)
Cohen, M. S. [147](#)
Cohen, P. [482](#), [521](#), [522](#), [523](#), [658](#), [668](#), [669](#), [690](#)
Cohen, S. [224](#), [230](#), [236](#), [240](#)
Cohn, J. [204](#)
Cohn, J. F. [367](#)
Cohn, L. [485](#)
Cohn, M. A. [87](#), [387](#)
Coie, J. D. [602](#)
Cole, D. A. [41](#), [581](#)
Cole, M. A. [581](#)
Cole, O. [598–599](#)
Cole, S. R. [49](#), [59](#)
Cole, S. W. [224](#), [230](#)
Coleman, M. R. [151](#)
Coleman, P. T. [262](#)
Collingwood, R. G. [11](#)
Collins, D. L. [132](#)
Collins, E. C. [267](#)
Collins, J. M. [60](#)
Collins, L. M. [552](#), [631](#), [638](#)
Collins, W. A. [380](#), [482](#)
Colman, A. M. [424](#)
Combs, D. [124](#)
Conaway, M. [425](#)
Condry, J. C. [191](#)
Conger, P. I. [534](#)

Conger, R. D. [534](#), [552](#), [589](#)
Conner, R. F. [53](#)
Conner, T. S. [231](#), [367](#), [373](#), [379](#), [384](#), [390](#)
Connolly, K. M. [223](#)
Connolly, T. [207](#)
Conover, W. J. [621](#)
Conrad, F. G. [425](#), [455](#)
Conrad, R. G. [424](#)
Conrath, D. W. [382](#)
Conrey, F. R. [201](#), [267](#)
Contractor, N. [201](#)
Converse, J. M. [422](#), [430](#), [431](#)
Converse, P. E. [425](#)
Cook, A. [379](#), [384](#)
Cook, C. [424](#), [458](#), [461](#)
Cook, J. [365](#)
Cook, R. D. [618](#), [619](#)
Cook, S. [83](#)
Cook, T. D. [11](#), [12](#), [16](#), [18](#), [21](#), [22](#), [27](#), [28](#), [35](#), [36](#), [39](#), [43](#), [45](#), [49](#), [50](#), [54](#), [59](#), [60](#),
[61](#), [62](#), [64](#), [65](#), [67](#), [69](#), [70](#), [73](#), [74](#), [88](#), [89](#), [90](#), [93](#), [395](#), [680](#)
Cook, W. L. [38](#), [362](#), [591](#), [592](#), [595–597](#), [603–604](#)
Cooley, P. [430](#)
Cools, R. [124](#)
Coombs, C. H. [425](#)
Coombs, L. C. [425](#)
Cooney, N. L. [389](#)
Cooper, D. R. [424](#)
Cooper, H. [677–678](#), [679](#), [681](#), [682](#), [689](#), [695–696](#), [697–698](#)
Cooper, J. [19](#), [663](#)
Cooper, W. H. [68](#)
Corneille, O. [660](#), [662](#), [671–672](#)
Cornell, D. P. [375](#), [379](#), [383](#)
Cornwell, B. [604](#)
Correll, J. [254](#), [276](#), [278](#), [602](#)
Cortina, J. M. [478](#), [689](#)
Cosgrove, G. R. [240](#)
Costa, P. T., Jr. [481](#), [482](#), [535](#)
Costafreda, S. G. [147](#)
Costin, F. [317](#)

Cotter, P. [431](#)
Cottrell, C. A. [2](#)
Couch, A. [424](#)
Coulter, P. [431](#)
Couper, M. P. [410](#), [415](#), [424](#), [425](#), [428](#), [429](#), [430](#), [433](#), [454](#), [455](#), [457](#), [461](#), [465](#),
[466](#)
Courchesne, E. [127](#)
Courneya, K. S. [423](#)
Court, C. A. [234](#)
Courvoisier, D. S. [392](#)
Covert, A. E. [142](#)
Cowan, W. [314](#)
Cox, C. J. [138](#)
Cox, D. R. [620](#)
Cox, G. M. [51](#)
Cox, R. W. [126](#)
Coye, R. W. [420](#)
Craig, S. B. [446](#), [456](#), [459](#)
Cramér, H. [357](#)
Crano, W. D. [11](#), [12](#), [14](#), [422](#)
Crawford, M. [118](#)
Crawford, S. D. [454](#), [457](#)
Crider, A. [102](#)
Critchley, H. D. [226](#), [229](#), [239](#)
Crites, S. L., Jr. [138](#), [139](#), [506](#)
Crockett, M. J. [225](#), [226](#), [227](#)
Cronbach, L. J. [22](#), [58](#), [82](#), [83](#), [360](#), [473](#), [477](#), [479](#), [480](#), [483](#), [485](#), [486](#), [688](#)
Crosby, F. [87](#)
Cross, T. [101](#)
Crow, M. D. [87](#), [347](#)
Cruise, L. J. [378](#)
Crutzen, R. [454](#)
Csikszentmihalyi, M. [4](#), [375](#), [379](#), [383](#), [384](#), [385](#)
Cubin, M. [176](#)
Cudeck, R. [170](#), [507](#), [508](#), [511](#), [512](#), [541](#)
Cuffin, B. N. [240](#)
Culhane, S. E. [474](#), [621](#), [667](#)
Cummings, A. [255](#)
Cummings, C. [445](#)

Cumsille, P. E. [631](#), [632](#)
Cunningham, W. A. [124](#), [135](#), [136](#), [137](#), [539](#)
Curhan, J. R. [599](#)
Curran, P. J. [513](#), [552](#), [668](#)
Curseu, P. L. [205](#)
Curtin, J. J. [142](#)
Curtin, M. [418](#)
Curtin, R. [418](#)
Cusimano, A. M. [597](#)
Custers, R. [324](#)
Cuthbert, B. N. [115](#), [118](#), [119](#), [220](#), [222](#), [226](#), [239](#), [240](#)
Czanner, S. [147](#)

D'Orofrío, B. M. [71](#)
D'Agostino, P. R. [321](#)
Dabbs, J. [87](#), [347](#)
Dabbs, J. M. [204](#), [207](#)
Dabbs, J. M., Jr. [589](#)
Dagne, G. [366–367](#)
Dahlke, A. [191](#)
Dalal, R. [446](#)
Dale, A. M. [126](#)
Dale, R. [254](#), [268](#), [269](#), [271](#), [274](#), [275](#), [276](#)
Dalkey, N. C. [210](#)
Dalrymple-Alford, E. C. [330](#)
Dalton, K. M. [143](#)
Daly, E. M. [226](#)
Damasio, A. R. [103](#)
Danforth, C. M. [270](#)
Dansereau, F. [598](#), [602](#)
Dapretto, M. [135](#), [150](#)
Darley, J. M. [190](#)
Darlington, R. B. [688](#)
Darwin, C. [116](#)
Das, M. [445](#), [462](#)
Dasen, P. R. [498](#)
David, J. P. [379](#), [384](#)
Davidenko, N. [446](#), [452](#), [456](#), [457](#)
Davidson, M. [148](#)

Davidson, R. J. [127](#), [143](#), [146](#), [147](#), [220](#), [221](#), [241](#)
Davidson, S. [176](#)
Davidson, W. A. [53](#)
Davies, P. G. [350](#)
Davies, T. [410](#)
Davis, A. [86](#), [373](#), [376](#), [390](#)
Davis, B. [360](#)
Davis, D. W. [431](#)
Davis, J. H. [190](#), [200](#)
Davis, J. I. [128](#), [224](#), [230](#)
Davis, K. E. [2](#)
Davis, M. [238](#)
Dawes, R. M. [191](#), [474–475](#), [489](#)
de Castro, C. C. [147](#)
De Dreu, C. K. W. [680](#)
de Geus, E. J. C. [178](#)
De Houwer, J. [228](#)
De Leeuw, E. D. [418](#)
De Leeuw, J. [57](#), [516](#), [517](#)
De Stavola, B. L. [164](#)
de Vries, M. W. [386](#)
de Weiger, A. D. [230](#)
Deacon, B. J. [684](#)
Dearing, M. F. [118](#)
Deaux, K. [380](#), [590](#)
DeBell, M. [413](#)
Decastro, J. M. [384](#)
Decety, J. [134](#), [231](#)
DeCharms, R. C. [230](#)
Deckman, T. [383](#)
Deckner, D. F. [270](#)
DeCoster, J. [267](#)
DeCourville, N. [381](#)
DeFour, D. [191](#)
DeFries, J. C. [161](#)
DeGelder, B. [241](#)
DeGraff, M. H. [425](#)
DeGuzman, C. [269](#)
Deichmann, R. [128](#), [130](#)

del Rosario, M. L. [65](#), [67](#)
Delaney, H. D. [32](#), [671](#)
Delbecq, A. L. [209](#), [210](#)
Delepaul, P. A. E. G. [383](#), [384](#), [386](#)
Delignières, D. [254](#), [269](#), [271](#), [275](#), [276](#)
DeLongis, A. [379](#)
Delplanque, S. [226](#)
DeMascio, A. [102](#)
Dempster, A. P. [627](#)
Deneubourg, J. L. [255](#)
Denissen, J. J. A. [159](#)
Dennis, A. R. [207](#)
Denny-Brown, C. [345](#)
DePaulo, B. M. [379](#), [383](#), [384](#)
Depue, B. E. [124](#), [148](#)
Derado, G. [134](#)
Des Jarlais, D. C. [430](#)
DeShon, R. P. [689](#)
DesJarlais, D. [73](#)
Desmond, J. E. [126](#)
DeSteno, D. A. [224](#), [230](#), [235](#), [515](#), [516](#), [520](#)
Deutsch, B. [423](#)
Deutsch, D. [316](#)
Deutsch, J. A. [316](#)
Deutsch, M. [191](#)
Devine, P. G. [137](#), [142](#), [326](#), [329](#), [335](#)
Devlin, J. T. [137](#)
Devos, T. [464](#)
DeVries, A. C. [230](#)
DeWall, C. N. [124](#), [379](#), [383](#), [690](#)
Dey, J. [446](#)
Dholakia, U. M. [445](#)
Diamond, L. M. [375](#)
Dickens, L. [230](#), [235](#)
Dickenson, J. [135](#)
Dickersin, K. [681](#)
Dickerson, S. [102](#), [103](#), [107](#), [118](#), [120](#)
Dickerson, S. S. [269](#)
Dickinson, A. [118](#)

Dickinson, T. L. [424](#)
Dickson, W. J. [194](#), [195](#)
Diehl, M. [193](#), [204](#), [207](#), [211](#)
Diener, E. [191](#), [346](#), [375](#), [378](#), [379](#), [381](#), [385](#), [544](#)
Dienstbier, R. A. [111](#), [112](#), [114](#)
Dieterich, M. [132](#)
Dietrich, D. [207](#)
DiGuseppi, C. [449](#), [454](#), [456](#), [457–458](#), [461](#), [464](#)
Dijksterhuis, A. [4](#), [324](#), [350](#)
Dill, C. A. [58](#), [60](#)
Dilla, W. N. [210](#)
Dillehay, R. C. [200](#)
Dillman, D. A. [417](#), [422–423](#), [429](#), [430](#), [433](#), [454](#), [455](#), [456](#), [461](#)
Dimberg, U. [227](#)
Dion, D. [267](#)
Dishion, T. J. [348](#), [353](#)
Dixit, J. [445](#)
Dixon, N. F. [313](#)
Dixon, W. J. [628–629](#)
Dockery, T. M. [604](#)
Dodds, P. S. [270](#)
Dodell-Feder, D. [146](#)
Dodge, K. A. [312](#), [602](#)
Doherty, T. [90](#)
Dolan, R. J. [135](#), [229](#)
Dollinger, S. J. [420](#)
Dolski, I. [143](#)
Donahue, E. M. [482](#), [542](#)
Dongxin, X. [269](#), [275](#)
Donnellan, M. B. [406](#), [408](#), [589](#), [590](#), [605](#)
Donner, A. [57](#)
Donnerstein, E. [21](#)
Donovan, C.-L. [147](#)
Donovan, J. E. [496](#)
Donovan, R. J. [425](#)
Dooley, K. J. [256](#)
Dorfman, D. [335](#)
Doss, R. C. [143](#)
Dougherty, D. [101](#)

Dougherty, F. E. [237](#)
Dovidio, J. F. [321](#), [332](#), [345](#)
Dowdney, L. [364](#)
Dowling, K. L. [210](#)
Downey, G. [376](#), [388](#), [394](#)
Downing, J. [425](#), [530](#)
Downs, D. [384](#)
Doyen, S. [318](#)
Doyle, M. C. [138](#)
Draine, S. C. [322](#)
Draper, D. [59](#)
Draper, N. R. [60](#)
Drasgow, F. [272](#)
Drastich, A. [147](#)
Dresel, C. [230](#)
Drigotas, S. M. [2](#)
du Toit, M. [560](#)
Duan, N. [73](#)
DuBois, R. M. [147](#)
Duchaine, B. C. [230](#)
Duck, S. [384](#)
Duckworth, K. L. [332](#)
Duclos, S. E. [224](#)
Duggan, E. W. [210](#)
Dumka, L. E. [54](#)
Duncan, S. [241](#), [242](#)
Duncker, K. [313](#)
Dunlap, W. P. [689](#)
Dunn, J. C. [33](#)
Dunton, B. C. [332](#)
Durantini, M. R. [690](#)
Duval, S. [693](#)
Dweck, C. S. [88](#)
Dwyer, T. [164](#)
Dyrenforth, P. S. [376](#)
Dzindolet, M. T. [207](#), [211](#)

Eagly, A. H. [318](#), [481](#), [677](#), [680](#), [681](#), [682](#), [683](#), [684](#), [687](#), [694](#), [698](#)
Eastwick, P. W. [599](#)

Ebbesen, E. B. [330](#)
Ebel, R. L. [425](#)
Ebersole, J. S. [137](#)
Ebner-Priemer, U. W. [395](#)
Ebstein, R. P. [176](#)
Eccles, J. [604](#)
Eckenrode, J. [380](#)
Eddy, J. M. [348](#), [353](#)
Eddy, W. F. [136](#)
Edelman, M. [418](#)
Edwards, A. W. F. [686](#)
Edwards, G. L. [225](#)
Edwards, J. H. [686](#)
Edwards, K. [321](#)
Edwards, M. C. [539](#), [553](#), [560](#), [561](#)
Edwards, P. [348](#), [449](#), [454](#), [455](#), [456](#), [457–458](#), [460](#), [461](#), [464](#)
Efron, B. [56](#)
Egger, M. [693](#)
Egloff, B. [599](#)
Eguíluz, V. M. [255](#)
Ehrhardt, A. A. [208](#)
Ehrlich, H. J. [425](#)
Ehrlich, P. R. [191](#)
Eich, E. [222](#), [228](#)
Eichele, T. [138](#)
Eickhoff, S. B. [144](#)
Eid, M. [346](#), [392](#), [544](#)
Eifermann, R. R. [424](#)
Eimer, M. [227](#), [241](#)
Eisdorfer, C. [677](#)
Eisemann, J. [73](#)
Eisenberger, N. I. [124](#), [134](#), [137](#), [147](#), [150](#), [225](#), [227](#), [384](#)
Eisenkraft, N. [599](#)
Ekkekakis, P. [224](#), [229](#)
Ekman, P. [116](#), [222](#), [234](#), [241](#)
Elfenbein, H. A. [599](#)
Elig, T. W. [423](#)
Ellenberg, J. [54](#)
Ellertson, N. [189](#), [202](#)

Elliot, A. J. [375](#)
Ellsworth, P. C. [31](#), [44](#), [83](#), [84](#), [242](#)
Eltinge, J. L. [418](#)
Embretson, S. E. [484](#), [485](#), [518–519](#), [566](#)
Emerson, J. D. [620](#)
Emerson, R. M. [85](#)
Emery, R. E. [162](#)
Emmelkamp, P. [225](#)
Emmons, R. A. [379](#), [380](#), [384](#), [385](#)
Enders, C. K. [53](#), [54](#), [627](#), [630](#), [632–634](#), [638](#), [640](#), [645](#), [646](#), [647](#)
England, L. R. [423](#)
Epley, N. [443](#), [447](#)
Epstein, D. H. [375](#)
Epstein, J. A. [379](#), [383](#), [384](#)
Epstein, L. H. [127](#)
Epstein, S. [380](#), [382](#), [391](#), [481](#)
Erb, M. [230](#)
Erber, R. [328](#), [336](#), [377](#)
Erdelyi, M. H. [313](#), [315](#)
Erdfelder, E. [53](#)
Erdle, S. [443](#)
Erdley, C. A. [321](#)
Ericsson, K. A. [322](#)
Ernst, C. [678](#)
Ernst, J. M. [113](#), [229](#), [240](#)

Essex, M. [664](#)
Ester, P. [462](#)
Estes, A. [353](#)
Etcoff, N. L. [227](#)
Ethier, N. [268](#), [272](#)
Etter, J. F. [447](#)
Evans, A. C. [132](#)
Evans, R. I. [58](#), [60](#)
Eveland, W. P., Jr. [408](#)
Everitt, B. S. [518](#)
Eyssell, K. M. [378](#)
Eyuboglu, N. [508](#)
Ezzyat, Y. [128](#)

Fabiani, M. [236](#)
Fabrigar, L. R. [33](#), [422](#), [505](#), [506](#), [507](#), [508](#), [509](#), [510](#), [520](#), [524](#), [527](#), [529](#), [535](#),
[538](#), [560](#)
Fahrenberg, J. [112](#), [115](#)
Fair, P. L. [237](#)
Fairbum, C. [176](#)
Fairchild, A. J. [44](#)
Fajen, B. R. [274](#)
Falaris, Evangelos M. [409](#)
Falk, E. B. [149](#), [150](#), [151](#), [376](#), [384](#)
Fan, W. [457](#), [458](#), [461](#), [464](#)
Farmer, T. A. [254](#), [268](#)
Farrington, D. P. [181](#)
Fast, K. [243](#)
Faucher, E. H. [444](#)
Faul, F. [53](#)
Faust, K. [202](#), [605](#)
Fayers, P. [560](#)
Fazio, R. H. [114](#), [316](#), [322](#), [323](#), [332](#), [333](#), [334](#), [520](#)
Fearn, M. [143](#)
Federmeier, K. D. [236](#)
Feinberg, T. E. [150](#)
Feldman, D. [56](#)
Feldman, J. [431](#)

Feldman, K. A. [6](#)
Feldman, L. A. [115](#), [147](#), [242](#), [376](#), [381](#), [515](#), [518](#)
Feldman, R. [269](#), [364](#)
Feldman, S. [337](#)
Feldner, M. T. [223](#)
Feldt, L. S. [35](#), [478](#)
Felleman, V. [684](#)
Felt, J. [445](#)
Fenaughty, A. M. [49](#), [74](#)
Fencsik, D. E. [138](#), [142](#)
Fennema-Notestine, C. [144](#)
Ferguson, G. A. [539](#)
Ferguson, M. J. [223](#), [311](#)
Fern, E. F. [208](#)
Fernández-Dols, J. M. [227](#), [237](#), [238](#)
Fernholz, L. T. [609](#)
Ferrer, E. [552](#)
Festinger, L. [85](#), [191](#), [196](#), [346](#)
Fetterman, J. D. [198](#)
Feuerstein, N. [314](#)
Fiddy, S. [176](#)
Fidell, L. S. [458](#), [459](#)
Fiedler, K. [581](#), [662](#)
Fiege, C. [60](#)
Figueredo, A. J. [484](#)
Filkins, J. [200](#)
Finch, J. F. [482](#), [491](#), [513](#)
Fincham, F. D. [383](#)
Fink, G. R. [127](#), [135](#)
Finkel, E. J. [379](#), [383](#), [599](#)
Finkel, S. E. [431](#)
Finney, J. W. [681](#)
Finzi, E. [230](#)
Fischer, B. [321](#)
Fischer, I. [211](#)
Fischer, R. [684](#), [690](#)
Fischl, B. [132](#)
Fishbein, M. [73](#), [481](#), [680](#)
Fisher, B. S. [430](#)

Fisher, J. D. [446](#)
Fisher, R. A. [57](#), [686](#), [690](#)
Fisk, D. W. [18](#)
Fiske, D. W. [4](#), [489](#), [534](#), [543](#)
Fiske, S. T. [1](#), [2](#), [85](#), [326](#), [327](#), [335](#), [381](#), [475](#), [525](#), [621](#)
Fitness, J. [381](#)
Fitzgerald, D. C. [379](#)
Fitzgerald, J. [409](#)
Fitzgerald, K. M. [445](#)
Fitzgerald, M. [136](#)
Fitzmaurice, G. M. [395](#)
Fitzpatrick, P. [269](#), [270](#)
Flack, W. F. [229](#)
Flagg, J. J. [198](#)
Flament, C. [191](#)
Flanagan, D. P. [534](#)
Flatt, M. [205](#)
Flax, R. [191](#)
Flay, B. R. [50](#)
Fleeson, W. [379](#), [380](#), [383](#)
Fleiss, J. L. [358–359](#), [484](#)
Fletcher, P. [136](#)
Flint, J. [176](#), [177](#)
Flores, S. A. [228](#), [229](#)
Floyd, F. J. [491](#), [538](#)
Foerde, K. [125](#)
Folkman, S. [379](#)
Fong, G. T. [44](#), [505](#), [663](#)
Fong, G. W. [134](#)
Fonov, V. [132](#)
Foos, P. W. [681](#)
Ford, B. Q. [228](#)
Forman, S. D. [136](#)
Förster, G. [384](#)
Forsyth, B. H. [428](#)
Forsyth, D. R. [188](#), [206](#)
Fortes, M. [254](#), [269](#), [271](#), [275](#), [276](#)
Foster, J. D. [690](#)
Fouraker, L. E. [191](#)

Fournier, M. A. [375](#), [384](#)
Fowler, C. A. [269](#), [271](#), [273](#), [274](#), [275](#), [322](#)
Fowler, F. J. [406](#), [417](#), [429](#), [431–432](#), [433](#)
Fowler, S. C. [116](#)
Fox, C. R. [134](#)
Fox, J. [148](#)
Fox, P. T. [144](#)
Fraley, R. C. [375](#), [383](#), [447](#), [448](#), [466](#)
Francis, G. [681](#)
Francis, M. E. [388](#)
Frangakis, C. E. [56](#)
Frankel, M. R. [414](#)
Frankel, S. [210](#)
Franklin, R. D. [65](#)
Franklin, S. G. [210](#)
Franks, N. R. [255](#)
Frederickson, B. L. [226](#)
Fredrickson, B. L. [378](#), [381](#)
Freedman, J. L. [191](#)
Freeman, J. B. [254](#), [267](#), [268](#)
Freeman, R. D. [128](#)
Freeman, W. H. [516](#), [518](#)
Freitas, A. L. [376](#), [394](#)
French, J. R. [190](#)
Frenkel-Brunswik, E. [424](#)
Fretz, R. I. [85](#)
Freund, P. A. [683](#)
Frey, J. H. [406](#), [429](#)
Fridlund, A. J. [116](#), [119](#), [234](#), [238](#)
Friedman, L. [147](#)
Friedman, M. P. [516](#), [518](#)
Friedrich, R. J. [669](#)
Friesen, W. V. [222](#), [234](#), [241](#)
Frieze, I. H. [423](#)
Friston, K. J. [123](#), [134](#), [135](#), [136](#)
Frith, C. D. [146](#), [147](#)
Frith, U. [146](#)
Fritz, M. S. [44](#), [658–659](#)
Fritzsche, B. A. [382](#)

Froming, W. J. [323](#)
Frost, J. H. [445](#), [447](#)
Fujita, F. [379](#)
Fulcher, R. [414](#)
Fuligni, A. J. [135](#)
Fulker, D. W. [178](#)
Fuller, J. L. [159](#)
Fuller-Rowell, T. [379](#), [380](#), [383](#)
Fullerton, J. [176](#)
Funder, D. C. [146](#), [378](#), [447](#)
Funk, J. L. [444](#)
Fury, J. M. [236](#)
Futoran, G. C. [198](#), [199](#)

Gable, S. L. [375](#), [376](#), [378](#), [380](#), [384](#), [388](#), [389](#), [391](#)
Gabrielli, J. D. E. [134](#), [146](#)
Gaffrey, M. S. [151](#)
Gagnon, J. H. [412](#)
Gaist, P. [73](#)
Galesic, M. [425](#)
Galili, G. [269](#)
Galindo, R. [74](#)
Gallagher, H. L. [146](#)
Gallagher, R. [254](#)
Gallhofer, I. N. [422](#)
Gallo, L. C. [380](#), [384](#)
Galván, A. [135](#)
Gamer, M. [230](#)
Gamez, W. [682](#)
Gamier, S. [269–270](#)
Ganellen, R. J. [323](#)
Gangadharan, S. P. [410](#)
Gapminder [684](#)
Garavan, H. [126](#)
Garcia, M. [332](#)
Garcia, R. L. [597](#), [599](#)
Garcia-Cueto, E. [424](#)
Gardner, R. G. [689](#)
Gardner, W. [365](#)

Gardner, W. L. [138](#), [139](#), [243](#)
Garrido, L. [230](#)
Gasking, D. [11](#)
Gaylord, R. H. [539](#)
Gebauer, J. E. [460](#)
Geen, T. R. [234](#)
Geer, J. G. [423](#)
Gehring, W. J. [138](#), [142](#)
Geister, S. [205](#)
Gelman, A. [37](#), [149](#)
Gendron, M. [227](#), [238](#), [242](#)
General Social Survey [684](#)
Genovese, C. R. [136](#)
Genshaft, J. L. [534](#)
Gentile, B. [690](#)
George, M. S. [230](#)
George, N. [241](#)
Georgiou, P. G. [367](#)
Gerard, H. B. [17](#), [18](#)
Gerber, A. S. [88](#), [90](#), [93](#), [414](#)
Gerber, J. C. [127](#)
Gerbing, D. W. [526](#)
German, E. [147](#)
Germán, M. [54](#)
Gerrard, M. [560](#), [563](#), [695](#)
Gerwien, D. P. [690](#), [695](#)
Ghiselli, E. E. [360](#)
Ghosh, C. M. [346](#), [352](#)
Gianaros, P. J. [239](#)
Gibbons, F. X. [695](#)
Gibbons, R. [539](#)
Gibbons, R. D. [560](#)
Gibson, K. [539](#)
Gigerenzer, G. [443](#)
Gil, K. M. [393–394](#), [525](#)
Gil, Y. [149](#)
Gilbert, C. A. [234](#)
Gilbert, D. [621](#)
Gilbert, D. T. [228](#), [326](#), [327](#), [328](#), [445](#), [475](#), [525](#)

Gilden, D. L. [275](#)
Gilkerson, J. [269](#), [275](#)
Gillan, G. [129](#)
Gillath, O. [383](#), [384](#)
Gilmour, R. [377](#)
Gilreath, T. D. [636](#), [637](#)
Ginexi, E. M. [56](#)
Gino, F. [581](#)
Ginsburg, G. P. [104](#), [105](#)
Glaser, J. [445](#)
Glaser, R. R. [60–61](#), [513](#)
Glasman, L. R. [690](#)
Glass, G. V. [681](#), [698](#)
Glass, L. [277](#)
Gleason, M. E. J. [379](#), [383](#)
Gleser, G. C. [360](#), [473](#), [480](#), [483](#)
Gleser, L. J. [695](#)
Glover, G. H. [125](#), [126](#), [147](#), [230](#), [231](#)
Gluck, M. A. [147](#)
Gnanadesikan, R. [610](#)
Gneezy, A. [88](#)
Gneezy, U. [88](#)
Gnisci, A. [364](#)
Gnys, M. [384](#)
Godbey, G. [375](#), [383](#), [385](#)
Goddard, M. E. [178](#)
Godwin, W. F. [198](#)
Goebel, R. [230](#)
Goel, S. [87](#)
Goense, J. B. M. [128](#)
Gofer, C. N. [314](#)
Goff, P. A. [350](#)
Goffman, E. [86](#), [115](#)
Goldberg, A. E. [591](#)
Goldberg, L. R. [475](#), [481](#), [482](#), [496](#), [497](#), [534](#), [535](#)
Goldberg, W. A. [684](#)
Goldberger, A. S. [64](#)
Goldman, R. [520](#)
Goldsmith, H. H. [220](#), [221](#)

Goldstein, A. G. [198](#)
Goldstone, R. L. [255](#)
Golembiewski, R. T. [209](#)
Gollob, H. F. [73](#)
Gollub, R. [147](#)
Gollwitzer, P. M. [317](#), [323](#), [324](#), [328](#), [335](#), [337](#)
Gonzales, M. H. [31](#), [44](#)
Gonzales, N. A. [54](#)
Gonzales, R. [37](#), [38](#)
Goodenough, F. L. [348](#)
Goodman, J. [273](#)
Goodstein, J. [149](#)
Goodwin, G. [176](#)
Goodwin, M. S. [65](#), [387](#)
Gorden, D. M. [255](#)
Gordis, E. [346](#), [352](#)
Gordon, S. [178](#)
Gordon, S. E. [330](#)
Göritz, A. S. [449](#), [452](#), [454](#), [464](#)
Gorman, B. S. [65](#)
Gorsuch, R. L. [510](#), [537](#)
Gosling, S. D. [28](#), [46](#), [74](#), [82](#), [85](#), [86](#), [87](#), [91](#), [160](#), [347](#), [443–444](#), [445](#), [446](#), [455](#),
[462](#), [463](#), [464](#)
Gotlib, I. H. [378](#)
Gottlieb, J. [254](#)
Gottman, J. M. [254](#), [258](#), [262](#), [272](#), [351](#), [352](#), [353](#), [354](#), [362](#), [363](#), [364](#), [365](#), [395](#)
Gottschalk, A. [254](#)
Gottschalk, P. [409](#)
Gottschall, A. [54](#)
Gottschall, A. C. [646](#)
Gough, H. G. [497](#)
Gould, S. J. [491](#)
Gould, William [409](#)
Govender, R. [332](#), [333](#)
Graesch, A. P. [347](#)
Gragani, A. [256](#)
Graham, F. K. [119](#)
Graham, J. [419](#)
Graham, J. W. [630](#), [631](#), [632–633](#), [636](#), [637](#), [640–644](#), [645](#), [647](#)

Granberg, D. [409](#)
Grand, S. [314](#)
Grandjean, D. [226](#)
Granger, R. L. [68](#)
Grant, A. M. [581](#)
Grant, J. T. [410](#)
Gratton, G. [236](#)
Gray, H. M. [345](#)
Gray, J. A. [111](#)
Gray, J. R. [146](#), [149](#)
Gray-Little, B. [480](#), [485](#)
Graziano, W. G. [51](#), [74](#), [84](#), [90](#), [384](#)
Green, A. S. [387](#)
Green, D. P. [27](#), [44](#), [88](#), [89](#), [90](#), [93](#), [414](#), [445](#), [664](#)
Green, M. C. [425](#), [431](#)
Green, P. J. [63](#), [379](#), [384](#)
Green, S. K. [677](#)
Greenacre, M. [518](#)
Greenberg, D. F. [408](#)
Greene, T. J. [210](#)
Greenholtz, J. [84](#)
Greenhouse, J. B. [687](#)
Greenwald, A. G. [19](#), [37](#), [314](#), [319](#), [322](#), [409](#), [445](#), [452](#), [466](#)
Greenwald, S. [680](#)
Greenwood, J. D. [24](#)
Gregory, D. [189](#), [202](#)
Grethe, J. S. [144](#)
Greve, D. [147](#)
Gribble, J. [430](#)
Grier, J. B. [330](#)
Griffin, D. [38](#)
Griffin, W. A. [365](#)
Griffiths, R. J. [328](#)
Grigg, J. A. [693](#), [694](#)
Grigoroudis, E. [206](#)
Grimm, O. [147](#)
Grimshaw, J. M. [697](#)
Grodd, W. [230](#)
Groenen, P. [515–516](#), [517–518](#)

Gross, J. J. [134](#), [146](#), [222](#), [238](#), [241](#), [242](#), [243](#), [269](#), [364](#), [487](#)
Grossman, P. [387](#)
Grove, J. B. [490](#)
Grove, W. M. [482](#)
Grover, J. B. [86](#), [87](#)
Groves, R. M. [381](#), [418](#), [429](#), [430](#), [433](#)
Grube, J. W. [423](#)
Grühn, D. [226](#)
Grundy, D. [246](#)
Guadagno, R. E. [445](#)
Guastello, S. J. [256](#), [257](#), [258](#), [265](#), [364](#)
Guest, L. [422](#)
Guethlein, W. [115](#)
Guetzkow, H. [202](#)
Guilford, J. P. [358](#), [537](#), [539](#)
Gulliksen, H. [476](#)
Gunderson, P. R. [364](#)
Gunthert, K. C. [385](#)
Guo, L. [415](#)
Guo, Y. [134](#)
Guo, Z. [415](#)
Gupta, A. [144](#)
Gureckis, T. M. [255](#)
Gurrin, L. C. [181](#)
Gurtman, M. B. [321](#), [332](#)
Gustafson, D. H. [209](#), [210](#)
Guterbock, T. M. [431](#)
Guthrie, D. [37](#)
Guthrie, T. J. [58](#), [60](#)
Gwet, K. [356](#), [358](#)

Ha, S. E. [44](#), [664](#)
Haas [352](#)
Haberstick, B. C. [162](#)
Hackman, J. R. [198](#)
Haddock, G. [422–423](#)
Hagger, M. S. [679](#)
Hahn, J. [63](#)
Haidt, J. [3](#), [419](#)

Hains, S. C. [192](#), [211](#), [602](#)
Hajcak, G. [230](#)
Haken, H. [273](#)
Hakstian, A. R. [507](#)
Halberstadt, J. B. [515](#), [516](#), [518](#)
Haley, E. [147](#)
Hall, D. L. [147](#)
Hall, J. A. [678](#)
Hallett, M. [123](#)
Hallmark, B. W. [605](#)
Halpern, C. T. [162](#)
Hamaker, E. L. [365](#)
Hamamura, T. [419](#)
Hambleton, R. K. [484](#), [553](#)
Hamer, D. [177](#)
Hamilton, D. L. [329](#), [330](#), [490](#)
Hamilton, W. [267](#)
Hampson, S. E. [475](#), [481](#)
Han, S. [419](#), [420](#)
Hancock, G. R. [547](#), [552](#)
Hancock, T. D. [480](#), [485](#)
Hanelin, J. [231](#)
Hanemann, W. M. [425](#)
Haney, C. [348](#)
Hanke, K. [690](#)
Hannaford, P. L. [197](#), [198](#)
Hänninen, L. [203](#)
Hannover, B. [322](#), [335](#)
Hans, V. P. [197](#), [198](#)
Hansen, D. [202](#)
Hansen, J. J. [464](#)
Hansen, K. M. [431](#)
Hansen, M. H. [432–433](#)
Hansson, K. [169](#)
Harbord, R. M. [694](#)
Harden, K. P. [162](#)
Hardy-Bayle, M. C. [134](#)
Hare, A. P. [188](#)
Hare, T. A. [128](#)

Haritos-Fatouros, M. [86](#)
Harkins, S. [190](#)
Harlow, L. [633–634](#)
Harlow, T. [375](#), [379](#), [383](#)
Harmann, S. [241](#)
Harmon-Jones, C. [143](#)
Harmon-Jones, E. [123](#), [127](#), [142](#), [143](#), [229](#), [236](#), [237](#), [241](#)
Harp, S. F. [149](#)
Harper, L. [270](#)
Harriea, C. [211](#)
Harris, C. [144](#)
Harris, G. T. [473](#)
Harris, K. D. [270](#)
Harris, K. M. [162](#)
Harris, M. J. [27](#), [346](#)
Harris, R. J. [37](#)
Harrison, P. L. [534](#)
Harrison, S. J. [271](#)
Harshman, R. A. [516](#), [517](#)
Hart, W. [694](#)
Hartley, S. L. [142](#)
Harvey, J. H. [188](#), [326](#)
Harvey, N. [211](#)
Harvey, O. J. [191](#), [194](#), [420](#)
Harzing, A.-W. [5](#), [423](#)
Haselgrove, C. [132](#)
Haseman, W. D. [206](#)
Hasher, L. [315](#)
Haslinger, B. [230](#)
Hasson, F. [210–211](#)
Hastie, R. [33](#), [38](#), [198](#), [199](#), [201](#), [327](#), [329](#)
Hatch, E. C. [207](#)
Hathaway, S. R. [497](#)
Hatsukami, D. K. [65](#)
Hattie, J. A. [495](#), [516](#), [517](#)
Haviland, A. [70](#)
Hawkley, L. C. [224](#), [230](#)
Haxby, J. [205](#)
Hayasaka, S. [136](#)

Hayes, A. F. [408](#), [523](#), [524](#), [527](#), [581](#), [659](#), [668](#), [674](#), [688](#)
Haynes, A. A. [408](#)
Haynes, S. N. [346](#), [352](#)
Hays, R. B. [379](#)
Hays, R. D. [560](#)
Haythorn, W. W. [188](#), [191](#)
Haythornthwaite, J. A. [390](#)
Hazzard, A. [349](#)
Heath, F. [424](#), [458](#), [461](#)
Heath, L. [65](#), [67](#)
Heath, R. A. [270](#), [272](#)
Heatherton, T. F. [137](#), [148](#)
Heckathorn, D. D. [421](#)
Heckhausen, H. [324](#)
Heckman, J. T. [647](#)
Hedges, L. V. [36](#), [45](#), [678](#), [685](#), [686](#), [687](#), [689](#), [692](#), [693](#), [694](#), [695](#), [697–698](#)
Hedges, S. M. [376](#), [383](#), [394](#)
Heider, F. [188](#)
Heine, S. J. [20](#), [84](#), [420](#), [422](#)
Heinsman, D. T. [681](#)
Heinze, H.-J. [119](#)
Heishman, S. J. [375](#)
Hektner, J. M. [373](#)
Helbing, D. [269–270](#)
Hellhammer, D. H. [224](#), [230](#)
Helmer, O. [210](#)
Helzer, J. E. [357](#)
Hemovich, V. [12](#)
Henchy, T. [74](#)
Henders, A. K. [178](#)
Henderson, A. H. [58](#), [60](#)
Henderson, A. J. Z. [200](#)
Hendin, H. M. [425](#), [480](#), [485](#)
Hendrick, C. [334](#)
Hendrie, C. A. [238](#)
Hendriksen, A. H. [203](#)
Henik, A. [656](#)
Hennenlotter, A. [230](#)
Hennigan, K. M. [65](#), [67](#)

Henrich, J. [20](#)
Henrich, T. R. [210](#)
Henry, G. T. [412](#), [413](#), [416](#)
Henry, P. J. [59](#), [83](#)
Henson, R. [136](#)
Henteleff, P. D. [71](#)
Hepworth, J. T. [67](#), [395](#)
Herd, S. A. [124](#), [148](#)
Herman, C. P. [330](#)
Hernandez, E. [603–604](#)
Herr, P. M. [322](#), [323](#)
Hertel, G. [205](#)
Hess, J. [427](#)
Hess, U. [40](#), [222](#), [227](#)
Hetenyi, M. A. [111](#), [114](#)
Hetherington, E. M. [168–169](#)
Hewitt, J. K. [178](#)
Hewstone, M. [328](#), [381](#)
Heyman, R. E. [348](#), [352–353](#), [361](#)
Hickcox, M. [378](#), [384](#)
Hicks, A. M. [375](#)
Higgins, C. A. [382](#)
Higgins, E. T. [38](#), [312](#), [313](#), [314](#), [317](#), [319](#), [322](#), [323](#), [335](#), [378](#)
Higgins, J. P. T. [692](#), [693](#), [694](#), [698](#)
Hill, A. [191](#)
Hill, G. W. [192](#)
Hill, J. [37](#), [149](#)
Hill, J. E. [162](#)
Hill, P. C. [58](#), [60](#)
Himmelfarb, S. [415](#), [474](#)
Himsel, A. [684](#)
Hinde, R. A. [375](#)
Hindy, C. G. [604](#)
Hinks, R. S. [147](#)
Hinsz, V. B. [200](#)
Hippler, H. J. [59](#), [423](#)
Hirano, K. [56](#)
Hirsch, H. R. [487](#)
Hirtz, D. [54](#)

Hitchcock, J. M. [238](#)
Hixon, J. G. [328](#)
Hluštík, P. [147](#)
Ho, D. E. [69](#)
Hoaglin, D. C. [620](#)
Hochberg, Y. [136](#)
Hodgson, D. [236](#)
Hoenig, K. [147](#)
Hoeppe, B. B. [65](#)
Hofer, S. M. [534](#), [631](#)
Hoffman, J. M. [658–659](#), [662](#)
Hoffman, L. R. [189](#)
Hoffman, R. [205](#)
Hofmann, W. [384](#)
Hofstede, G. [684](#)
Hogan, H. [489–490](#)
Hogan, R. [482](#)
Hogg, M. A. [192](#), [211](#), [602](#)
Hojatkashani, C. [132](#)
Holbert, R. L. [527](#), [581](#)
Holbrock, J. [238](#)
Holbrook A. L. [418](#), [425](#), [431](#)
Holden, J. G. [271](#), [275](#), [276](#)
Holden, J. H. [256](#)
Holender, D. [322](#)
Holland, C. [12](#)
Holland, P. W. [49](#), [50](#), [56](#)
Holley, W. [358](#)
Hollingshead, A. B. [205](#), [206](#)
Hollman, M. [151](#)
Holmberg, S. [409](#)
Holmes, A. [227](#), [241](#)
Holmes, J. G. [83](#), [85](#), [87](#), [380](#)
Holt, R. [190](#)
Holtgrave, D. [73](#)
Holton, A. [348](#)
Hommer, D. [134](#)
Hon, A. [334](#)
Hong, G. [57](#)

Hong, S. [509](#), [510](#), [515](#), [538](#)
Hood, D. [53](#)
Hood, W. [191](#), [194](#)
Hoover, J. B. [432](#)
Hopewell, S. [681](#), [683](#)
Hopko, D. R. [223](#)
Hopper, C. H. [207](#)
Hopper, J. L. [164](#)
Hoppmann, C. [597](#)
Hops, H. [348](#), [360](#)
Hori, I. [419](#)
Hori, K. [238](#)
Hormuth, S. E. [379](#), [384](#)
Horn, E. M. [605](#)
Horn, J. [552](#)
Horn, J. L. [534](#), [547](#), [548](#)
Horrigan, J. [446](#)
Hothersall, D. [427](#)
Hottenga, J. [178](#)
Housts, A. C. [59](#), [74](#)
Houts, A. C. [395](#)
Houts, C. R. [561](#)
Hovland, C. I. [420](#), [505](#)
Howe, G. W. [366–367](#)
Howell, R. D. [526](#)
Hox, J. J. [57](#), [575](#), [576](#), [592](#)
Hoyle, R. H. [27](#), [513](#), [660](#)
Hoyt, C. L. [231](#)
Hromi, A. [377](#), [384](#)
Hsiung, S. [62](#), [64](#)
Hsu, L. M. [443](#)
Hu, L. [513](#), [541](#)
Hu, X. [60–61](#)
Huang, C. [231](#), [240](#)
Huang, J. Y. [200](#)
Huang, X. [599](#), [602](#)
Hubbard, J. A. [602](#)
Huber, G. A. [91](#)
Huedo-Medina, T. B. [684](#), [690](#), [693](#), [694](#), [695](#)

Huettel, S. A. [123](#)
Hufford, M. R. [373](#), [387](#), [389](#)
Hugdahl, K. [138](#)
Huggins, M. K. [86](#)
Hughes, B. [134](#)
Hughes, J. N. [70](#)
Hughes, W. W. [86](#), [87](#)
Hull, J. G. [494](#)
Hummel, T. [127](#)
Humphreys, L. G. [508](#)
Hunt, M. [697](#)
Hunt, S. M. J. [205](#)
Hunter, J. E. [487](#), [689](#), [698](#)
Hunter, J. S. [334](#)
Hunter, S. B. [115](#), [350](#)
Hunter, W. G. [334](#)
Hurd, A. W. [422–423](#)
Hurd, M. D. [423](#)
Hurst, M. H. [384](#)
Hutchins, D. [208](#), [209](#)
Hutton, C. [128](#), [130](#)
Hyde, J. S. [147](#)
Hyers, L. L. [379](#)
Hyman, H. [200](#), [421](#), [431](#)
Hymes, C. [332](#), [333](#)

Iacono, W. G. [170](#)
Ialongo, N. S. [56](#)
Iberall, A. [255](#)
Ickes, W. I. [326](#), [377](#)
Iida, M. [379](#), [383](#)
Ijumba, P. [90](#)
Ijzerman, H. [446–447](#)
Imai, K. [60](#), [69](#)
Iman, R. L. [621](#)
Imbens, G. W. [49](#), [55](#), [56](#), [63](#), [64](#), [70](#)
Inagaki, T. K. [134](#), [226](#)
Ingle, S. [209](#)
Innes-Ker, A. [229](#)

Insightful Corp. [366](#)
Insko, C. A. [376](#)
International Labor Organization [684](#)
International Social Survey Programme [684](#)
International Telecommunication Union [443](#), [466](#)
Intille, S. S. [384](#), [387](#)
Ioannidis, J. P. A. [681](#), [693](#)
Ipeirotis, P. G. [463](#)
Irwin, J. R. [670](#)
Irwin, M. R. [134](#)
Irwin, W. [147](#)
Isenhower, R. [255](#), [273](#)
Ito, T. A. [117](#), [139](#), [226](#), [232](#), [233](#), [240](#)
Iwata, J. [238](#)
Iyengar, S. [687](#)
Izard, C. E. [234](#), [237](#)
Izuma, K. [124](#)

Jabbi, M. [225](#)
Jaccard, J. [34](#), [665](#)
Jackman, M. R. [424](#)
Jackson, D. [90](#)
Jackson, D. C. [143](#)
Jackson, D. J. [423](#)
Jackson, D. N. [535](#)
Jackson, J. E. [424](#)
Jackson, J. R. [332](#)
Jacobson, G. C. [410](#)
Jacobson, L. [663](#)
Jacobson, N. S. [361](#)
Jacoby, J. [424](#), [663](#)
Jacoby, L. L. [45](#), [336](#)
James, J. [229](#)
James, L. R. [407](#), [654](#)
James, T. W. [223](#)
James, W. [85](#), [220](#), [315](#), [337](#)
Jamieson, J. [238](#)
Jandorf, L. [383](#)
Jansen, H. [203](#)

Jansen, R. G. [203](#)
Janssen, L. [350](#)
Japec, L. [418](#)
Jarvis, B. [205](#)
Jasechko, J. [336](#)
Javitz, H. S. [416](#), [421](#), [431](#)
Jeffress, L. A. [313](#)
Jeffries, E. [230](#)
Jelièiæ, H. [627](#)
Jenike, M. A. [227](#)
Jenkins, G. M. [67](#)
Jenkins, J. G. [422](#)
Jenkinson, M. [131](#)
Jensen-Campbell, L. A. [384](#)
Jesmanowicz, A. [147](#)
Jessor, R. [496](#)
Jetten, J. [490](#)
Jeuniaux, P. [269](#)
Jian, J.-Y. [210](#)
Jiang, X. L. [598](#)
Jo, B. [56](#)
Joahansen-Berg, H. [131](#)
Jobe, J. B. [380](#)
Johannesen-Schmidt, M. C. [684](#)
John, L. K. [622](#)
John, O. P. [128](#), [160](#), [443–444](#), [445](#), [446](#), [464](#), [481](#), [482](#), [484](#), [485](#), [487](#), [489–490](#),
[493](#), [495](#), [496](#), [497](#), [499](#), [504](#), [542](#)
John, R. S. [394](#)
Johnson, B. T. [678](#), [680](#), [681](#), [682](#), [683](#), [684](#), [685](#), [689](#), [690](#), [691](#), [692](#), [694](#), [695](#)
Johnson, C. [207](#)
Johnson, H. H. [190](#), [198](#)
Johnson, J. A. [91](#), [446](#), [455](#), [456](#), [458](#)
Johnson, J. C. [560](#)
Johnson, J. T. [421](#)
Johnson, M. K. [315](#)
Johnston, D. [128](#)
Johnstone, T. [228](#)
Jolesz, F. A. [230](#)
Jolly, J. P. [423](#)

Jones, C. R. [314](#), [319](#), [322](#)
Jones, E. E. [2](#), [101](#), [110](#)
Jones, L. E. [34](#), [669](#)
Jones, L. W. [423](#)
Jordon, T. [66](#), [67](#)
Jöreskog, K. G. [479](#), [536](#), [541](#), [544](#), [547](#)
Joseph, J. [228](#), [241](#)
Josephs, O. [128](#), [130](#), [147](#)
Jovicich, J. [147](#)
Joyce, C. A. [141](#)
Judd, C. M. [27](#), [34](#), [35](#), [37](#), [38](#), [42](#), [43](#), [65](#), [67](#), [191](#), [194](#), [409](#), [421](#), [474](#), [475](#), [484](#),
[486](#), [491](#), [496](#), [505](#), [520](#), [522](#), [524](#), [528](#), [529](#), [589](#), [602](#), [615](#), [620](#), [621](#), [623](#),
[654](#), [656](#), [657](#), [659](#), [660](#), [662](#), [665](#), [667](#), [668](#), [669–670](#), [673–674](#)
Judice, T. N. [494](#)
Juechems, K. [445](#)
Julien, D. [364](#)
Juslin, P. N. [228](#)
Jussim, L. [604](#)

Kaczmirek, L. [451](#)
Kafer, N. F. [201](#)
Kagan, J. [382](#), [493](#)
Kahn, R. L. [200](#)
Kahn, R. S. [147](#)
Kahneman, D. [2](#), [378](#), [381](#), [383](#), [475](#)
Kalar, D. [144](#), [148](#), [149](#)
Kalton, G. [412](#), [413](#)
Kaltz, D. [114](#)
Kam, C. D. [408](#)
Kam, C. M. [638](#)
Kamarck, T. W. [111](#)
Kamaya, K. [419](#)
Kamber, M. [132](#)
Kameda, T. [201](#)
Kamin, L. J. [180](#)
Kan, L. P. [229](#)
Kanade, T. [367](#)
Kang, L. [69](#)
Kang, Y. [146](#)

Kanouse, D. E. [409](#), [424](#)
Kantowitz, B. H. [328](#)
Kantz, H. [271](#), [272](#)
Kaplan, D. [277](#)
Kaplan, E. H. [90](#)
Karasawa, M. [419](#)
Karau, S. J. [687](#)
Kardes, F. R. [316](#), [332](#), [333](#)
Karelina, K. [230](#)
Kärnä, A. [586](#)
Kashy, D. A. [38](#), [362](#), [379](#), [383](#), [384](#), [491](#), [512](#), [525](#), [589](#), [590](#), [591](#), [592](#), [595–597](#), [598](#), [600](#), [603–604](#), [605](#)
Kasimatis, M. [375](#), [383](#), [394](#)
Kasprowicz, A. L. [111](#)
Kasprowicz, A. S. [111](#)
Kassel, J. A. [384](#)
Kasten, N. [683](#)
Katigbak, M. S. [498](#)
Katkin, E. S. [111](#), [114](#)
Katsamanis, A. [367](#)
Katz, D. [421](#)
Katz, L. B. [324](#), [329](#), [330](#)
Katz, M. [176](#)
Katz, N. [201](#)
Kauff, D. M. [321](#)
Kazdin, A. E. [67](#)
Keele, L. [60](#)
Keeney, S. [210–211](#)
Keeter, S. [418](#)
Keil, F. C. [149](#)
Kelderman, H. [485](#)
Kelley, C. [134–135](#), [336](#)
Kelley, H. H. [83](#), [85](#), [87](#), [191](#), [375](#), [393–394](#)
Kello, C. T. [256](#)
Kellstedt, P. M. [408](#)
Kelly [360](#)
Kelly, A. M. [134–135](#), [148](#)
Kelly, J. R. [198](#), [199](#)
Kelsey, R. M. [104](#), [112](#), [115](#)

Kelso, J. A. S. [254](#), [256](#), [257](#), [263](#), [269](#), [273](#)
Keltner, D. [222](#), [227](#), [235](#), [242](#), [243](#)
Kemeny, M. E. [107](#), [269](#)
Keniston, K. [424](#)
Kennedy, D. P. [127](#)
Kennerley, A. J. [128](#)
Kenny, D. A. [14](#), [35](#), [37](#), [38](#), [39](#), [40](#), [44](#), [60](#), [65](#), [67](#), [350](#), [362](#), [393–394](#), [407](#), [484](#),
[491](#), [519](#), [520](#), [523](#), [525](#), [528](#), [544](#), [580](#), [589](#), [592](#), [595–597](#), [598](#), [599](#), [600](#),
[602](#), [603–604](#), [605](#), [654](#), [656](#), [657](#), [658](#), [660](#), [662](#), [664](#)
Kent, R. N. [198](#)
Kentle, R. L. [482](#), [542](#)
Keppel, G. [27](#)
Keren, G. [272](#)
Kerig, P. K. [351](#)
Kern, R. P. [226](#)
Kernis, M. H. [375](#), [379](#), [383](#)
Kerr, N. L. [83](#), [85](#), [87](#), [190](#), [192](#), [198](#), [199](#), [200](#), [209](#), [210](#)
Kessler, R. C. [379](#), [390](#), [408](#)
Kessous, L. [270](#)
Keysers, C. [225](#)
Khoury, H. [376](#), [394](#)
Khuder, S. A. [66](#), [67](#)
Kibler, J. L. [113](#), [229](#), [240](#)
Kidd, R. F. [326](#)
Kidder, L. H. [194](#)
Kiecolt-Glaser, J. K. [525](#)
Kiernan, M. [664](#)
Kiesler, S. [205](#), [206](#)
Kihlstrom, J. F. [323](#)
Kilgore, K. [348](#)
Kilkowski, J. M. [475](#)
Killingsworth, M. A. [228](#)
Killworth, P. [381](#)
Kim, A. S. N. [228](#)
Kim, C. [513](#), [525](#)
Kim, H. S. [115](#), [116](#), [233](#)
Kim, M. J. [227](#)
Kim, S.-H. [413](#)
Kimmel, H. D. [499](#)

Kincses, Z. T. [147](#)
Kinder, D. R. [407–408](#), [409](#), [410–411](#)
King, G. [53](#), [69](#)
King, G. A. [38](#), [323](#)
King, J. E. [484](#)
King, L. A. [384](#)
Kinnear, T. C. [423](#)
Kirk, R. E. [27](#), [39](#)
Kirkendol, S. E. [379](#), [383](#), [384](#)
Kirkman, B. L. [684](#)
Kirsch, I. [684](#)
Kirschbaum, C. [224](#), [230](#)
Kirsner, K. [33](#)
Kirson, D. [115](#)
Kish, L. [59](#), [412](#), [413](#), [414](#), [417](#)
Kitayama, S. [20](#), [419](#)
Kite, M. E. [685](#), [689](#)
Kittur, A. [144](#), [149](#)
Kivlighan, D. R. [598–599](#)
Klar, N. [57](#)
Klasmeyer, G. [228](#)
Klein, K. J. [192](#)
Klein, O. [318](#)
Kleinberger, E. [211](#)
Kleiner, B. [610](#)
Kleiner, M. [146](#)
Klerman, G. L. [237](#)
Klevansky, S. [191](#)
Kline, R. B. [576](#), [586](#)
Klinger, M. R. [319](#)
Klockars, A. J. [424](#)
Klohn, E. C. [487](#)
Klonsky, B. G. [682](#)
Kloos, H. [256](#), [275](#)
Klorman, R. [229](#)
Kloumann, I. M. [270](#)
Klumb, P. [597](#)
Knäuper, B. [423](#)
Knee, C. R. [384](#)

Knierim, K. [231](#)
Kniesner, Thomas J. [409](#)
Knight, R. T. [119](#), [128](#)
Knoke, D. [202](#)
Knonacki, R. [60–61](#)
Knowlton, B. J. [125](#)
Knutson, B. [134](#)
Kober, H. [128](#), [147](#), [227](#), [228](#), [232](#), [241](#), [242](#)
Koch, A. E. [236](#)
Koenig, A. M. [680](#), [687](#)
Koenig, L. B. [161](#)
Koestler, A. [311](#)
Koestner, R. [381](#)
Koffka, K. [312](#)
Kogan, N. [30](#)
Kogan, W. [190](#)
Kohler, H. [409](#)
Kohut, A. [418](#)
Kolodziej, M. E. [695](#)
Kolstein, R. [265](#)
Komorita, S. S. [191](#)
Konorski, J. [118](#)
Konradt, U. [205](#)
Konrath, S. [690](#)
Koo, M. [445](#)
Koola, J. [230](#)
Koopman, R. F. [507](#)
Kopans, D. B. [53](#)
Kopp, R. J. [425](#)
Korchmaros, J. D. [393–394](#)
Korhonen, L. J. [210](#)
Koriat, A. [314](#)
Korteweg, T. [147](#)
Kotov, R. [682](#)
Kotsch, N. E. [237](#)
Kowai-Bell, N. [115](#), [350](#)
Kozlowski, S. W. J. [192](#)
Kraemer, H. C. [664](#)
Kramer, G. P. [200](#)

Kramer, R. M. [191](#), [204](#)
Kramer, T. J. [207](#)
Krantz, D. H. [516](#), [518](#)
Krantz, D. S. [111](#)
Krantz, J. H. [445](#), [446](#), [460](#)
Kratochwill, T. R. [67](#)
Krauss, R. M. [191](#)
Kraut, R. E. [409](#), [465](#), [466](#)
Kravitz, D. A. [206](#)
Kreft, I. G. G. [57](#)
Kreibig, S. D. [240](#)
Kreuter, F. [430](#)
Kreuz, R. J. [270](#)
Kriegeskorte, N. [144](#)
Kring, A. M. [234](#)
Kringelback, M. L. [241](#)
Kroehne, U. [60](#)
Krokoff, L. J. [352](#)
Kromrey, J. D. [682](#)
Kronauer, R. E. [269](#)
Kronenfeld, D. [381](#)
Krosnick, J. A. [406](#), [407–409](#), [412](#), [413](#), [416](#), [418](#), [421](#), [422](#), [423](#), [424](#), [425](#), [431](#),
[529](#), [530](#)
Kross, E. [148](#)
Krueger, A. B. [383](#)
Krueger, R. [170](#)
Krueger, R. A. [208](#)
Krueger, R. F. [161](#)
Krueger, A. [88](#)
Kruger, J. [443](#), [447](#)
Kruglanski, A. W. [337](#), [378](#), [420–421](#)
Krull, D. S. [326](#), [327](#), [328](#)
Krull, J. L. [660](#), [661](#), [662](#)
Kruskal, J. B. [515](#), [516–517](#)
Kuhl, C. K. [147](#)
Kuklinski, J. H. [202](#)
Kulik, J. [23](#)
Kumar, P. A. [33](#), [38](#), [327](#)
Kuo, W.-L. [552](#)

Kupfer, D. J. [664](#)
Kuppens, P. [225](#), [375](#), [379](#)
Kurth, F. [144](#)
Kurtzman, H. S. [380](#)
Kutas, M. [241](#)
Kwak, N. [408](#)
Kwan, I. [449](#), [454](#), [455](#), [456](#), [457–458](#), [461](#), [464](#)
Kwang, T. [462](#), [463](#)
Kwok, O.-M. [57](#)

Lac, A. [12](#)
Lacefield, K. [690](#)
LaFrance, M. [83](#)
Lagarde, J. [269](#)
Lago, P. P. [210](#)
Lahey, B. B. [71](#)
Laird, A. R. [144](#)
Laird, J. D. [224](#), [229](#)
Laird, N. M. [395](#), [627](#)
Lakenberg, N. [178](#)
LaLonde, F. [205](#)
Lam, C. [599](#), [602](#)
Lambert, N. M. [383](#)
Lambie, J. A. [242](#)
Lambon Ralph, M. A. [230](#)
Lamias, M. J. [410](#), [454](#), [457](#)
Lamm, H. [30](#), [41](#)
Lammert, A. C. [367](#)
Lampkin, E. C. [198](#)
Landau, M. J. [445](#)
Landau, S. [518](#)
Lane, R. D. [384](#)
Lang, A.-G. [53](#)
Lang, K. [226](#)
Lang, P. J. [115](#), [118](#), [119](#), [220](#), [221](#), [222](#), [223](#), [226](#), [227](#), [228](#), [237](#), [239](#), [240](#)
Langer, E. J. [149](#), [326](#), [337](#)
Langlois, J. H. [382](#)
Lareau, A. [86](#)
Larkin, G. R. [226](#)

Larrance, D. T. [229](#)
Larsen, J. T. [139](#), [232](#), [233](#), [240](#)
Larsen, R. J. [375](#), [379](#), [383](#), [385](#), [394](#)
Larson, J. R. [201](#)
Larson, R. W. [375](#), [377](#), [383](#), [384](#)
Lashley, K. S. [313](#)
Lasswell, H. D. [102](#)
Latané, B. [190](#)
Lau, J. [693](#)
Laughlin, P. R. [190](#), [198](#), [207](#)
Laukka, P. [228](#)
Laumann, E. O. [412](#)
Laurenceau, J.-P. [373](#), [379](#), [390](#), [393](#), [394](#), [577](#), [605](#)
Laurent, A. [430](#)
Lautenschlager, G. J. [446](#)
Lavrakas, P. J. [406](#), [417](#), [418](#), [429](#), [431](#), [432](#), [433](#)
Law, H. G. [516](#), [517](#)
Lawrence, D. M. [383](#)
Lawrence, F. R. [552](#)
Lawrennce, E. [446](#)

Lazar, N. A. [136](#)
Lazarus, R. S. [379](#)
Lazer, D. [201](#)
Lazkowski, D. K. [598](#)
le Roux, I. [90](#)
Le, B. [457](#)
Le, H. [689](#)
Le, K. [376](#)
Leaf, P. J. [49](#), [59](#)
Leahy, R. M. [240](#)
Learner, S. M. [383](#)
Leary, M. R. [384](#), [460](#)
Leavitt, H. J. [189](#)
Leblanc, A. [605](#)
Ledermann, T. [595–597](#), [605](#)
Ledgerwood, A. [659–660](#)
LeDoux, J. E. [125](#), [238](#)
Lee, C. C. [367](#)
Lee, H. K. [419–420](#)
Lee, I. [678](#), [682](#)
Lee, J. [684](#)
Lee, M. B. [227](#)
Lee, S. [445](#)
Lee, S. H. [178](#)
Lee, Y. [54](#)
Lee-Chai, A. Y. [323](#)
Leese, M. [518](#)
LeFevre, J. [4](#)
Légaré, F. [605](#)
Leggett-Dugosh, K. [207](#)
Lehman, B. J. [147](#)
Lehman, D. R. [68](#), [84](#), [419](#), [420](#), [422](#)
Lehman, J. [563](#)
Lehmann, D. [137](#)
Lehn, D. A. [494](#)
Leigh, B. C. [375](#)
Leinhardt, S. [427](#), [658](#)
Leirer, V. O. [324](#), [330](#)

Leitten, C. L. [112](#)
Leivers, S. [425](#)
Lelkes, Y. [412](#)
Lemay, E. P., Jr. [589](#)
Lemieux, T. [63](#), [64](#)
Lemoine, L. [275](#)
Lempert, R. O. [68](#)
Lench, H. C. [228](#), [229](#)
Lenhart, A. [446](#)
Lennon, C. A. [690](#), [695](#)
Lennox, R. [232](#)
Lenz, G. S. [91](#)
Leon, D. A. [164](#)
Leong, F. T. [420](#)
Lepkowski, J. M. [418](#), [429](#), [433](#)
Lerner, J. S. [235](#), [242](#), [243](#)
Lerner, M. [327](#), [329](#)
Lerner, M. J. [114](#)
Lerner, R. M. [627](#)
LeScanff, C. [225](#)
Lesgold, A. M. [330](#)
Leslie, W. D. [145](#)
Lessler, J. T. [428](#)
Leung, K. [498](#)
Levendusky, M. S. [416](#), [421](#), [431](#)
Levenson, R. W. [222](#), [269](#), [364](#), [365](#)
Levesque, M. J. [350](#)
Levin M. L. [201](#)
Levin, J. R. [67](#)
Levine, J. M. [188](#), [205](#)
Levine, S. [375](#), [379](#), [385](#)
Levinson, D. J. [424](#)
Levitt, M. [681](#)
Levy, R. [547](#)
Lewenstein, M. [266](#)
Lewin, K. [81](#), [83](#), [190](#), [253](#)
Lewine, J. D. [240](#)
Lewis, C. [541](#)
Lewis, P. A. [229](#)

Lewontin, R. C. [180](#)
Li, J. [204](#)
Li, N. P. [683](#), [689](#)
Li, X. [415](#)
Li, Y. [445](#)
Liamputtong, P. [208](#)
Liang, K.-Y. [362](#)
Liberati, A. [683](#)
Libkuman, T. M. [226](#)
Lichtenstein, M. [330](#)
Lichtenstein, P. [169](#)
Lickel, B. [115](#), [350](#)
Lieberman, M. D. [123](#), [124](#), [135](#), [136](#), [144](#), [146](#), [147](#), [150](#), [151](#), [225](#), [226](#), [227](#),
[376](#), [384](#)
Liebovitch, L. S. [258](#), [275](#)
Lien, J. [367](#)
Light, R. J. [693](#)
Liker, J. K. [363](#)
Lillard, Lee [409](#)
Lilley, C. M. [234](#)
Lin, I. F. [419](#)
Lin, Y. [384](#)
Linacre, J. M. [560](#)
Lind, J. C. [541](#)
Lindahl, K. M. [351](#)
Lindberg, C. M. [226](#)
Lindberg, M. J. [694](#)
Lindenberger, U. [538](#)
Linder, D. E. [86](#)
Lindner, N. M. [464](#)
Lindquist, K. A. [147](#), [221](#), [226](#), [227](#), [228](#), [232](#), [235](#), [236](#), [238](#), [240](#), [241](#), [242](#), [243](#)
Lindquist, M. A. [147](#)
Lindsay, D. S. [336](#)
Lindzey, G. E. [201](#), [314](#), [422](#), [474–475](#), [525](#), [621](#)
Link, M. W. [414](#), [418](#)
Linn, R. L. [507](#)
Lipka, R. P. [481](#)
Lipkus, I. [517](#)
Lippett, R. [190](#)

Lipsey, M. W. [681](#), [682](#), [683](#), [689](#), [690](#), [691](#), [698](#)
Lischetzke, T. [392](#), [544](#)
Litt, M. D. [389](#), [390](#)
Little, R. J. A. [53](#), [418](#), [627](#), [628](#), [632–633](#)
Little, T. D. [538](#), [539](#), [547](#), [576](#), [579](#), [580](#), [582](#), [584](#), [586](#)
Liu, T. J. [319](#)
Livi, S. [598](#), [602](#)
Lockwood, C. M. [658](#), [659](#), [662](#)
Lockwood, P. [375](#), [384](#)
Lockyer, J. [605](#)
Lodewijkx, H. F. M. [207](#), [211](#)
Lodge, M. [424](#)
Loehlin, J. C. [172](#), [491](#)
Loehr, D. [270](#)
Loevinger, J. [485](#)
Loewenstein, G. [622](#)
Logan, G. D. [314](#), [327](#), [331](#), [335](#), [336](#)
Logothetis, N. K. [128](#)
Lohr, S. [59](#)
Lomax, R. G. [512](#), [525](#), [527](#)
Lombardi, W. J. [34](#), [316](#), [320–321](#), [323](#)
London, E. D. [135](#)
Long, J. S. [491](#), [512](#), [541](#)
Longoria, N. [360](#)
Loomis, J. [231](#)
Loosveldt, G. [431](#)
Lorber, M. F. [365](#)
Lorch, R. F., Jr. [333](#)
Lord, F. [476](#), [478](#), [482](#), [484](#), [565](#)
Lorenz, K. [346](#)
Losch, M. E. [115](#), [116](#), [233](#)
Lotze, M. [230](#)
Loucks, R. A. [227](#)
Louwerse, M. M. [269](#)
Louzoun, Y. [269](#)
Lovallo, W. R. [112](#), [115](#), [225](#)
Low, C. A. [236](#)
Lozano, L. M. [424](#)
Luby, J. L. [151](#)

Lucas, R. E. [376](#), [379](#), [406](#)

Lucas-Thompson, R. [684](#)

Luce, R. D. [516](#), [518](#)

Luck, S. J. [123](#), [139](#)

Ludlow, D. H. [230](#), [231](#)

Lüdtke, O. [586](#), [587](#)

Ludwig, J. [62](#), [423](#)

Lueck, L. [143](#)

Luedtke, O. [539](#)

Lui, J. [204](#)

Luke, D. A. [53](#)

Luke, J. V. [419](#)

Luminet, O. [660](#), [662](#)

Lumsden, J. [273](#)

Lunn, D. J. [560](#)

Luo, S. [507](#)

Lusk, R. [191](#)

Luskin, R. C. [425](#)

Lutsky, N. [677](#)

Lutz, C. [220](#), [238](#)

Lützkendorf, R. [151](#)

Lyberg, L. [428](#)

Lykken, D. T. [159](#)

Lynn, S. K. [226](#)

Ma, D. [690](#)

MacCallum, R. C. [505](#), [508](#), [509](#), [510](#), [513](#), [514](#), [515](#), [516](#), [517](#), [518](#), [525](#), [527](#),
[529](#), [538](#), [539](#), [542](#), [582](#), [585](#), [671](#)

MacCoun, R. [198](#), [200](#), [209](#)

MacDonald, D. [132](#)

MacDonald, G. [124](#)

MacDonald, J. [205](#)

Macfarlan, S. J. [683](#)

Machilek, F. [445](#)

Macho, S. [597](#)

Mackenbach, J. P. [447](#)

Mackey, S. C. [230](#), [231](#)

Mackie, J. L. [11](#)

MacKinnon, D. P. [44](#), [59](#), [60](#), [522](#), [631](#), [657](#), [658–659](#), [660](#), [661](#), [662](#)

MacLeod, C. [335](#)
MacLin, M. K. [203](#)
MacLin, O. H. [203](#)
Macrae, C. N. [124](#), [125](#), [273](#), [327](#), [328](#), [490](#)
MacWhinney, B. [205](#)
Madansky, A. [610](#), [613](#), [621](#)
Madden, M. [446](#)
Madhyastha, T. M. [365](#)
Madow, W. G. [432–433](#)
Maechler, M. [586](#)
Maeda, F. [230](#)
Maes, H. [586](#)
Magidson, J. [68](#)
Maglio, S. J. [226](#)
Magori-Cohen, R. [269](#)
Maher, M. P. [417](#)
Maier, N. R. F. [189](#)
Maital, N. [227](#)
Mak, Y. E. [127](#)
Makhijani, M. G. [682](#), [687](#)
Malafosse, A. [447](#)
Malarkey, W. B. [230](#), [525](#)
Malave, V. L. [151](#)
Malisza, K. L. [145](#)
Malkoff, S. B. [111](#)
Malle, B. F. [238](#), [680](#)
Maluccio, J. A. [409](#)
Mandel, M. R. [237](#)
Mandelbrot, B. B. [275](#)
Mangione, T. W. [429](#), [431–432](#)
Mangun, G. R. [119](#)
Maniaci, M. R. [445](#), [446](#), [456](#), [457](#), [459](#), [463](#)
Maniewski, R. [240](#)
Mann, C. B. [409](#)
Mann, V. [164](#)
Manne, S. [603–604](#)
Mannetti, L. [598](#)
Mantovani, F. [225](#), [231](#)
Manuck, S. B. [111](#)

Maoz, B. [176](#)
Mar, R. A. [147](#), [228](#)
Marcel, A. J. [242](#), [320](#)
Marchand, A. [225](#)
Marcia, J. [200](#)
Marco, C. [384](#)
Marco, C. A. [378](#)
Marcus, B. [445](#)
Marcus, D. K. [603–604](#), [605](#)
Marcus-Newhall, A. [684](#)
Marecek, R. [147](#)
Margolin, G. [346](#), [352](#), [362](#), [363](#), [394](#)
Marien, H. [324](#)
Marín-Martínez, F. [692](#), [693](#)
Maringer, M. [227](#)
Mark, A. L. [239](#)
Mark, M. M. [62](#), [69](#)
Markon, K. E. [170](#)
Marks, D. F. [228](#)
Marks, E. S. [421](#)
Marks, M. M. [325](#)
Markus, H. [333](#)
Markus, H. R. [20](#), [419](#)
Marquette, J. [418](#)
Marquis, K. H. [430](#)
Marrone, G. F. [375](#)
Marschat, L. E. [422–423](#)
Marsden, P. V. [417](#)
Marsh, H. W. [481](#), [539](#), [541](#), [544](#), [573](#), [586](#), [587](#)
Marsh, K. L. [255](#), [269](#), [273](#), [276](#)
Marshall-Goodell, B. [116](#)
Marsland, A. L. [236](#)
Martel, M. O. [225](#)
Martens, A. [444](#)
Martin, A. [124](#)
Martin, B. [206](#)
Martin, E. [428](#)
Martin, J. [384](#), [415](#), [428](#)
Martin, L. L. [229](#)

Martin, R. [379](#), [384](#)
Martindale, J. [128](#)
Martzke, J. S. [233](#)
Marwan, N. [274](#)
Maser, J. [497](#)
Mashal, N. M. [134](#)
Mashek, D. [457](#)
Mason, R. [547](#), [548](#)
Mason, R. A. [151](#)
Mason, W. [87](#)
Mason, W. A. [267](#)
Masten, C. L. [124](#), [135](#), [137](#), [150](#)
Masters, G. N. [558](#)
Masterson, F. A. [118](#)
Matell, M. S. [424](#)
Mathalon, D. H. [147](#)
Mathes, J. [668](#)
Mathewson, G. C. [17](#), [18](#)
Mathiak, K. [230](#)
Mathiowetz, N. [428](#)
Matsatsinis, N. F. [206](#)
Matsumoto, D. [222](#), [445](#)
Matt, G. E. [60](#), [74](#)
Matthews, K. A. [380](#), [384](#)
Mauricio, A. M. [54](#)
Mauss, I. B. [232](#), [269](#), [364](#)
Mavin, G. H. [38](#)
Mavros, P. L. [146](#)
Maxwell, S. E. [32](#), [41](#), [58](#), [60](#), [581](#), [671](#)
Mayer, A. [60](#)
Mayer, R. E. [149](#)
Mayhew, J. E. [128](#)
Mayo, E. [195](#)
Mayr, U. [375](#)
Mazur-Hart, S. F. [64](#)
Mazzulla, E. C. [227](#)
McAdams, D. P. [374](#)
McArdle, J. J. [165](#), [491](#), [493](#), [496](#), [534](#), [547](#), [548](#), [552](#)
McArthur, D. [356](#), [360](#)

McArthur, L. Z. [2](#)
McBride, D. [189](#), [202](#)
McBride, L. [198](#)
McBride, M. [422–423](#)
McCabe, D. P. [149](#)
McCabe, P. M. [272](#)
McCall, M. A. [67](#)
McCarter, L. [269](#), [364](#)
McCarthy, D. M. [487](#)
McCarthy, G. [123](#)
McCarthy, P. J. [421](#)
McCaulley, M. H. [497](#)
McClain, D. B. [54](#)
McClean, R. J. [382](#)
McClelland, D. C. [375](#), [381](#)
McClelland, G. H. [34](#), [35](#), [38](#), [42](#), [43](#), [474](#), [475](#), [484](#), [486](#), [491](#), [522](#), [528](#), [615](#),
[620](#), [621](#), [623](#), [659](#), [660](#), [662](#), [665](#), [666](#), [667](#), [668](#), [669–670](#)
McClelland, J. L. [267](#)
McClendon, M. J. [425](#)
McConahay, J. B. [409](#)
McConnell, H. K. [427](#)
McCormick, R. A. [537](#)
McCrae, R. R. [481](#), [482](#), [535](#)
McCulloch, W. [255](#)
McDonald, R. P. [516](#), [517](#), [539](#), [541](#), [560](#), [561](#)
McDonel, E. C. [114](#), [323](#)
McEvoy, B. P. [178](#)
McFarland, C. [381](#)
McFarland, S. G. [426](#)
McGarva, A. [269](#)
McGaugh, J. L. [238](#)
McGaw, B. [681](#), [698](#)
McGhee, D. E. [445](#), [452](#)
McGinnies, E. [313](#)
McGlennen, K. M. [135](#)
McGrath, J. E. [188](#), [191](#), [192](#), [193](#), [194](#), [195](#), [197](#), [198](#), [199](#), [205](#), [206](#), [211](#), [256](#)
McGrath, P. J. [234](#)
McGrath, R. E. [696](#)
McGraw, K. [348](#)

McGraw, K. O. [359](#), [484](#), [696](#)
McGue, M. [161](#), [170](#)
McGuire, S. [168–169](#)
McGuire, T. [205](#)
McGuire, W. J. [2](#), [3](#), [81](#), [82](#), [84](#), [86](#), [94](#), [376](#)
McInemey, S. C. [227](#)
McInnis, M. G. [236](#)
McIntosh, D. N. [229](#)
McIntyre, S. H. [423](#)
McKenna, C. [348](#)
McKenna, H. [210–211](#)
McKinley, J. C. [497](#)
McKinstry, R. C. [132](#)
McMullen, A. [384](#)
McNaughton, N. [111](#)
McNealy, K. [150](#)
McNeil, D. W. [223](#)
McNulty, J. K. [379](#), [383](#), [571](#), [574–575](#), [577](#)
McRae, K. [134](#)
McReynolds, P. [496](#)
McSweeny, A. J. [53](#)
McTavish, J. [191](#)
Mead, G. H. [253](#)
Mead, M. [346](#)
Meade, A. W. [446](#), [456](#), [459](#), [463](#)
Meadows, D. H. [258](#)
Mechelli, A. [123](#)
Medina, A. M. [346](#), [352](#)
Meehl, P. E. [253](#), [485](#), [486](#)
Meek, D. [190](#)
Meeren, H. K. M. [241](#)
Mehl, M. R. [87](#), [90](#), [231](#), [347](#), [367](#), [373](#), [375](#), [379](#), [383](#), [384](#), [387](#), [390](#), [604](#)
Mehta, M. A. [147](#), [149](#)
Meier, S. C. [445](#)
Meijer, Z. [419](#)
Meinke, A. [274](#)
Meiser, T. [581](#), [662](#)
Melinat, E. [84](#)
Mellenbergh, G. J. [484](#)

Mellor, S. [62](#)
Mels, G. [507](#)
Mendelsohn, G. A. [221](#), [237](#)
Mendes, W. B. [102](#), [103](#), [107](#), [115](#), [118](#), [120](#), [224](#), [238](#), [242](#), [345](#), [350–351](#)
Mendle, J. [162](#)
Mendoza-Denton, R. [376](#), [383](#)
Menon, G. [382](#)
Meredith, W. [547](#), [548](#), [552](#)
Mergeche, J. [269](#)
Mericer, M. [241](#)
Merkle, D. [418](#)
Mermillod, M. [227](#)
Merrill, L. [694](#)
Mesquita, B. [227](#), [238](#), [241](#), [242](#), [243](#)
Messick, D. M. [204](#)
Messick, S. [487](#), [488](#), [491](#)
Metcalf, J. [222](#), [228](#)
Meuwese, J. D. [230](#)
Meyer, D. E. [331](#), [332](#)
Meyer, E. S. [203](#)
Meyer, G. J. [696](#)
Meyvis, T. [446](#), [452](#), [456](#), [457](#)
Michael, J. [204](#)
Michael, R. T. [412](#)
Michaelis, B. [376](#), [394](#)
Michaels, S. [412](#)
Michaelson, L. K. [210](#)
Michalyszyn, D. [225](#)
Michel, C. M. [137](#)
Michels, L. C. [446](#)
Middeldorp, C. M. [178](#)
Middleton, R. T. [413](#)
Mier, D. [147](#)
Miethe, T. D. [423](#), [424](#)
Mikels, J. A. [223](#), [226](#)
Mikl, M. [147](#)
Mikulincer, M. [2](#), [383](#), [384](#)
Miles, D. R. [161](#)
Miles, L. K. [273](#)

Milgram, S. [11](#), [20](#), [190](#)
Milham, M. P. [134–135](#)
Miller, A. G. [44](#)
Miller, C. [418](#)
Miller, D. L. [62](#)
Miller, D. T. [696](#)
Miller, E. [144](#), [149](#)
Miller, E. K. [148](#)
Miller, G. [384](#), [387](#), [395](#), [445](#)
Miller, G. F. [159](#)
Miller, J. J. [39](#)
Miller, J. M. [425](#)
Miller, L. C. [254](#)
Miller, M. B. [147](#), [148](#)
Miller, N. [117](#), [684](#), [696](#)
Miller, N. H. [114](#)
Miller, P. [427](#)
Miller, S. [176](#)
Milleville, S. C. [124](#)
Mills, J. [17](#), [659](#), [663](#)
Millsap, R. E. [54](#), [539](#), [547](#)
Milne, A. B. [327](#), [490](#)
Milz, S. [66](#), [67](#)
Minbashian, A. [376](#)
Minder, C. [693](#)
Ming, K. [70](#)
Minor, J. K. [229](#)
Mintun, M. A. [136](#)
Mirman, D. [275](#)
Mirowsky, J. [425](#)
Mirza, M. [132](#)
Mischel, W. [254](#), [312](#), [337](#), [374](#), [380](#), [490](#)
Mishler, E. G. [202](#)
Mislevy, R. J. [560](#)
Mitchell, A. A. [680](#), [687](#)
Mitchell, J. P. [124](#), [125](#), [151](#)
Mitchell, M. A. [581](#)
Mitchell, R. C. [425](#)
Mitchell, S. [355](#), [360](#)

Mitra, S. [270](#)
Mittleman, B. [102](#)
Miyake, K. [375](#), [384](#)
Mize, J. [241](#), [242](#)
Mobley, M. F. [424](#)
Moffitt, R. [409](#)
Moffitt, T. E. [161](#), [490](#)
Moher, D. [683](#)
Mohr, C. D. [350](#), [377](#), [384](#)
Mokdad, A. H. [414](#)
Molenaar, P. C. [67](#)
Molnar, C. [230](#)
Monahan, J. L. [326](#)
Moneta, G. B. [379](#)
Montanelli, R. G., Jr. [508](#)
Monterosso, J. [128](#)
Montgomery, C. J. [234](#)
Montgomery, G. W. [177](#), [178](#)
Monti, M. M. [134](#), [151](#)
Moody, W. R. [425](#)
Mook, D. G. [21](#), [22](#), [31](#)
Mookerjee, A. [445](#)
Mooney, M. E. [65](#)
Moore, B. [23](#)
Moore, D. W. [425](#), [426](#)
Moore, L. R. [177](#)
Moore, T. J. [684](#)
Moore, V. M. [164](#)
Moors, A. [228](#)
Moosman, M. [138](#)
Moran, J. M. [148](#)
Morel, S. [241](#)
Moreland, R. L. [188](#), [189](#), [192](#), [198](#), [211](#), [598](#), [602](#)
Moreno, J. L. [194](#), [201](#)
Morf, C. C. [444](#)
Morgan, B. [230](#)
Morgan, S. L. [69](#)
Morganti, F. [225](#), [231](#)
Morgen, K. [147](#)

Morgenthau, S. [609](#)
Morin, A. J. S. [539](#)
Morland, L. [346](#), [352](#)
Morley, R. [164](#)
Morris, C. G. [198](#)
Morris, J. [135](#)
Morris, J. S. [230](#)
Morris, S. B. [689](#)
Morrison, S. F. [239](#)
Morse, P. [389](#)
Mortensen, C. R. [377](#)
Moser, S. E. [70](#)
Mosher, J. C. [240](#)
Moskowitz, D. S. [254](#), [384](#)
Moskowitz, G. B. [312](#), [324](#), [328](#), [334](#), [335](#), [337](#)
Moss, A. J. [384](#)
Moss, W. B. [457](#)
Mosteller, F. [421](#), [620](#)
Mott, R. [176](#)
Mourão, M. L. [147](#)
Moussaïd, M. [269–270](#)
Mowbray, C. T. [53](#)
Moyer, A. [681](#)
Mueller, C. J. [143](#)
Muellerleile, P. A. [680](#)
Mugridge, C. [111](#), [114](#)
Mulac, A. [207](#)
Mullen, B. [207](#), [684](#)
Mullen, P. [73](#)
Müller, C. [151](#)
Muller, D. [520](#), [524](#), [655](#), [659](#), [673–674](#)
Müller, M. I. [387](#)
Mumford, J. A. [123](#), [126](#), [130](#)
Munafo, M. R. [160](#), [177](#)
Munakata, Y. [124](#), [148](#)
Muñiz, J. [424](#)
Munson, J. M. [423](#)
Munsterman, G. T. [197](#), [198](#)
Münste, T. F. [119](#), [241](#)

Muraki, E. [539](#), [558–559](#), [560](#)
Muraven, M. B. [329](#)
Murphy, G. M. [226](#)
Murphy, K. [126](#), [147](#)
Murphy, M. D. [330](#)
Murphy, S. A. [60](#)
Murphy, S. T. [326](#)
Murray, D. M. [57](#), [446](#)
Murray, J. [258](#), [262](#), [365](#)
Muthén, B. O. [162](#), [170](#), [364](#), [366–367](#), [512](#), [539](#), [552](#), [560](#), [563](#), [573](#), [586](#), [587](#)
Muthén, L. [162](#), [170](#)
Muthén, L. K. [364](#), [367](#), [552](#), [560](#)
Myers, D. G. [30](#), [41](#)
Myers, I. B. [497](#)
Myers, J. H. [424](#)
Myers, T. A. [523](#), [527](#), [581](#)

N'diaye, K. [226](#)
Nagao, D. H. [200](#)
Nagengast, B. [60](#), [539](#), [586](#)
Nagin, D. S. [70](#)
Nagy, V. T. [409](#)
Nahas, Z. [230](#)
Nair, V. N. [60](#)
Nalivaiko, E. [239](#)
Nanda, H. [360](#), [473](#), [483](#)
Narayanan, S. S. [367](#)
Narr, K. L. [132](#)
Nathan, B. R. [423](#)
Nathan, L. R. [118](#)
National Telecommunications and Information Administration [446](#)
Nayak, D. A. [223](#)
Neale, J. M. [375](#), [376](#), [378](#), [379](#), [381](#), [383](#), [384](#), [388](#), [394](#)
Neale, M. [586](#)
Neale, M. C. [176](#)
Nebeling, L. [381](#)
Neberich, W. [460](#)
Necowitz, L. B. [514](#)
Neely, J. H. [316](#), [331–332](#)

Neiderhiser, J. M. [161](#), [169](#)
Neidhart, E. [447](#)
Neisser, U. [316](#)
Nelson, C. A. [147](#)
Nelson, D. [427](#)
Nelson, K. [54](#)
Nelson, L. [88](#)
Nelson, L. D. [443](#), [622](#)
Nemanov, L. [176](#)
Nesselrode, J. R. [375](#), [515](#), [516](#), [517](#), [538](#)
Netemeyer, R. G. [424](#)
Neubarth, W. [461](#)
Neuberg, S. E. [326](#)
Neuberg, S. L. [2](#), [323](#), [494](#)
Neuhaus, C. [449](#)
Neumann, O. [314](#)
Newcomb, T. M. [189](#)
Newell, A. [315](#)
Newman, J. C. [430](#)
Newman, K. B. [348](#)
Newman, L. S. [322](#)
Newsom, J. T. [49](#), [74](#)
Newston, D. [276](#)
Nezlek, J. B. [375](#), [379](#), [384](#)
Ng, Z. W. [447](#)
Nichols, T. E. [123](#), [126](#), [130](#), [136](#), [144](#), [147](#)
Nichols, W. L. [415](#)
Niedenthal, P. M. [227](#), [229](#), [321](#), [515](#), [516](#), [518](#)
Nielsen-Bohlman, L. [119](#)
Nielson, R. [53](#)
Nijstad, B. A. [207](#), [211](#), [680](#)
Nils, F. [222](#)
Ninot, G. [254](#), [269](#), [271](#), [275](#), [276](#)
Nisbett, R. E. [68](#), [151](#), [224](#), [230](#), [322](#), [333](#), [419](#), [427](#), [696](#)
Nitschke, J. B. [143](#)
Nitz, W. [129](#)
Noelle-Neumann, E. [423](#)
Noftle, E. E. [380](#)
Noguchi, K. [690](#)

Noland, B. [586](#)
Noldus, L. P. [203](#)
Nolen-Hoeksema, S. [690](#)
Noll, D. C. [136](#)
Noll, J. [534](#)
Noma, E. [201](#)
Nordby, H. [138](#)
Norenzaya, A. [20](#)
Norman, C. [348](#)
Norman, D. A. [316](#)
Norman, G. J. [224](#), [230](#)
Norris, F. H. [415](#)
Norris, K. [111](#), [114](#)
Northoff, G. [150](#)
Northway, M. L. [201](#)
Norton, M. I. [445](#), [447](#)
Nosek, B. A. [419](#), [445](#), [452](#), [464](#), [466](#)
Notarius, C. I. [364](#)
Novak, N. [226](#)
Novick, M. R. [476](#), [478](#), [482](#)
Novick, O. [176](#)
Nowak, A. [254](#), [257](#), [259](#), [262](#), [263](#), [265](#), [266](#), [267](#), [276](#), [277](#)
Nunnally, J. C. [480](#)
Nussbeck, F. W. [544](#)
Nystrom, L. W. [146](#)

O'Brien, B. [273](#)
O'Brien, W. H. [346](#), [352](#)
O'Conner, C. [115](#)
O'Connor, K. [417](#)
O'Connor, K. M. [205](#)
O'Connor, S. C. [375](#)
O'Craven, K. M. [126](#), [147](#)
O'Grady, E. [446](#)
O'Hare, B. [417](#)
O'Hearn, H. G. [346](#), [352](#)
O'Neil, K. M. [453](#), [458](#), [466](#)
O'Reilly, J. M. [415](#)
O'Reilly, R. C. [124](#), [148](#)

O'Dell, L. L. [511](#), [512](#)
Oaten, M. J. [236](#)
Obrist, P. A. [110](#)
Ochs, E. [347](#)
Ochs, L. [53](#)
Ochsner, K. N. [128](#), [134](#), [146](#), [148](#), [224](#), [230](#), [231](#), [238](#), [241](#), [242](#), [243](#)
Office of Information and Regulatory Affairs [418](#)
Offner, A. K. [207](#)
Oh, I. [689](#)
Oh, I.-S. [689](#)
Olatunji, B. O. [223](#)
Olchowski, A. E. [631](#), [636](#), [637](#)
Olekalns, M. [592](#), [598](#)
Oler, J. A. [227](#)
Olinsky, A. [633–634](#)
Oliver, P. [346](#), [352](#)
Olkin, I. [36](#), [45](#), [678](#), [692](#), [693](#), [694](#), [695](#), [697–698](#)
Oller, D. K. [269](#), [275](#)
Olsen, M. K. [638](#)
Olson, G. [204](#)
Olson, I. R. [128](#)
Olson, J. [465](#), [466](#)
Olson, J. S. [204](#)
Oltmanns, T. F. [159](#)
Ong, A. D. [379](#), [380](#), [383](#)
Onizuka, R. K. [517](#)
Oosterhof, N. N. [2](#)
Oppenheimer, D. M. [446](#), [452](#), [456](#), [457](#)
Oravec, Z. [375](#), [379](#)
Orlando, M. [560](#)
Orne, M. [12](#), [17](#), [32](#)
Orpaz, A. [383](#), [384](#)
Ortony, A. [220](#), [243](#)
Orwin, R. G. [684](#), [693](#)
Osborn, A. F. [206](#), [207](#)
Osborn, L. [414](#)
Osborne, R. E. [326](#), [327](#)
Osburn, H. G. [209](#)
Osgood, C. E. [228](#)

Osher, Y. [176](#)
Osherson, D. N. [149](#)
Oskenberg, L. [427](#), [430](#)
Ostrom, T. M. [330](#)
Otani, H. [226](#)
Otter-Henderson, K. D. [375](#)
Over, H. [324](#)
Owen, A. M. [124](#)
Owens, J. A. [164](#)
Owren, M. J. [222](#), [227](#), [228](#), [234](#), [237](#)
Oxley, N. L. [207](#)
Ozer, D. J. [489](#)

Page-Gould, E. [376](#), [383](#)
Palermo, R. [241](#)
Paluck, E. L. [27](#), [85](#), [88](#), [89](#)
Panksepp, J. [220](#)
Panter, A. T. [444](#)
Paolacci, G. [463](#)
Paone, D. [430](#)
Papadakis, N. [128](#)
Parasuraman, R. [205](#)
Pardo, S. T. [445](#)
Parfitt, G. [224](#), [229](#)
Park, B. [191](#), [324](#), [329](#), [602](#)
Park, C. L. [453](#)
Park, H. [419](#)
Park, J. [209](#)
Park, N. [444](#)
Park, S. [209](#)
Parker, D. S. [148](#)
Parker, J. [443](#), [447](#)
Parkinson, B. [381](#)
Partchev, I. [60](#)
Pascual-Marqui, R. D. [137](#)
Pasek, J. [412](#), [413](#)
Pashler, H. [144](#)
Pastell, M. [203](#)
Pasupathi, M. [375](#)

Patall, E. A. [695–696](#)
Patching, G. R. [138](#)
Patel, N. [11](#)
Patel, S. [227](#), [237](#)
Paton, J. [239](#)
Patrick, D. L. [424](#)
Patterson, G. R. [347–348](#), [351](#), [353](#)
Paty, J. A. [382](#), [384](#)
Paul, B. Y. [117](#)
Paul, S. [206](#)
Paulhus, D. L. [490](#)
Paulus, M. P. [145](#)
Paulus, P. B. [207](#), [211](#)
Pauly, J. M. [230](#)
Pavlicová, M. [147](#)
Payne, B. K. [412](#)
Payne, E. [177](#)
Payne, K. [412](#)
Payne, S. L. [425](#)
Pearcey, S. M. [384](#)
Pearl, J. [69](#)
Pearlman, K. [487](#)
Pearrow, M. J. [146](#)
Pearson, J. E. [446](#)
Pearson, K. [686](#)
Pedersen, W. C. [684](#)
Pederson, N. L. [169](#), [178](#)
Pedhazur, E. J. [480](#)
Pedley, T. A. [137](#)
Pelham, B. W. [326](#), [327](#), [328](#)
Pemberton, M. B. [376](#)
Peng, K. [84](#)
Penke, L. [159](#)
Pennebaker, J. W. [87](#), [90](#), [347](#), [375](#), [387](#), [388](#)
Penner, L. A. [345](#), [382](#)
Pennington, N. [198](#), [199](#)
Penrod, S. D. [198](#), [199](#), [453](#), [455](#), [458](#), [466](#)
Pequegnat, W. [73](#)
Perdue, C. W. D. [321](#), [332](#)

Pereira, G. M. [209](#)
Perozo, N. [269–270](#)
Perry-Jenkins, M. [377](#)
Pervin, L. A. [482](#), [504](#)
Peters, H. E. [409](#)
Petersen, C. K. [127](#)
Petersen, S. E. [126](#)
Peterson, C. [444](#)
Peterson, C. K. [229](#), [236](#)
Peterson, D. A. M. [408](#), [411](#)
Peterson, D. R. [384](#)
Peterson, R. A. [444](#)
Peterson, R. S. [411](#)
Petridou, N. [147](#)
Petruzzello, S. J. [224](#), [229](#)
Pettigrew, T. F. [682](#)
Petty, R. E. [19](#), [21](#), [33](#), [41](#), [102](#), [115](#), [116](#), [119](#), [233](#), [322](#), [405](#), [425](#), [487–488](#),
[506](#), [520](#), [524](#), [530](#), [581](#), [655](#), [656](#)
Peugh, J. L. [627](#)
Pew Internet and American Life Project [443](#), [446](#)
Pfeifer, J. H. [135](#), [150](#), [225](#), [227](#)
Pfeiffer, U. [127](#)
Pfent, A. M. [418](#)
Phelps, E. [627](#)
Philippot, P. [222](#)
Philipsen, G. [207](#)
Phillips, M. A. [57](#), [74](#)
Phillips, M. L. [226](#)
Pichon, C.-L. [318](#)
Pickard, J. D. [151](#)
Pickering, A. [239](#)
Pickles, A. R. [364](#)
Pierce, K. [127](#)
Pierro, A. [598](#)
Pietromonaco, P. R. [317](#), [319](#), [321](#), [322](#), [376](#), [381](#)
Pigott, T. D. [682](#)
Pike, A. [168–169](#)
Piliavin, J. A. [345](#)
Pilkington, C. J. [589](#)

Pillemer, D. B. [693](#)
Pincomb, G. A. [225](#)
Pincus, D. [257](#), [364](#)
Ping, R. A. [528](#)
Pinkus, R. T. [375](#), [384](#)
Pinto, A. [353](#)
Pirke, K. M. [224](#), [230](#)
Pitcher, D. [230](#)
Pittman, T. [322](#)
Pizzagalli, D. [236](#)
Pleyers, G. [660](#), [662](#)
Plichta, M. M. [147](#)
Plomin, R. [161](#), [168–169](#)
Plonsey, R. [137](#)
Plotzker, A. [128](#)
Pobric, G. [230](#)
Poehlmann, K. M. [232](#), [233](#), [240](#)
Pohl, S. [73](#)
Poirier-Bisson, J. [225](#)
Poldrack, R. A. [123](#), [124](#), [125](#), [130](#), [134](#), [135](#), [136](#), [137](#), [143](#), [144](#), [147](#), [149](#)
Poline, J.-B. [147](#)
Pond, R. S. [379](#), [383](#)
Ponz, A. [241](#)
Poortinga, Y. H. [498](#)
Pope, J. E. [53](#)
Porges, S. W. [272](#)
Pornprasertmanit, S. [582](#)
Port, R. F. [276](#)
Portela, L. A. P. [147](#)
Posner, M. I. [315](#), [316](#), [331](#), [336](#)
Post, D. L. [2](#)
Postman, L. [313](#)
Potenza, M. N. [128](#)
Potter, J. [443](#), [444](#), [445](#), [464](#)
Powell, C. [124](#)
Powell, L. [146](#)
Powell, M. C. [316](#), [332](#), [333](#)
Powers, E. A. [677](#)
Powers, M. B. [225](#)

Pratap, S. [449](#), [454](#), [455](#), [456](#), [457–458](#), [461](#), [464](#)
Prather, A. A. [236](#)
Pratto, F. [332](#), [333](#), [335](#), [678](#), [682](#)
Pratto, P. [332](#)
Prause, J. [684](#)
Preacher, K. J. [393–394](#), [507](#), [510](#), [523](#), [524](#), [525](#), [527](#), [538](#), [580](#), [581](#), [582](#), [585](#),
[587](#), [655](#), [656](#), [662](#), [668](#), [671](#), [674](#)
Preiser, N. [577](#)
Prelec, D. [622](#)
Prentice, D. A. [696](#)
Prescott, C. A. [165](#)
Prescott, S. [383](#)
Press, J. [176](#)
Presser, S. [410](#), [418](#), [422](#), [425](#), [426](#), [428](#), [430](#)
Preston, C. C. [375](#)
Preston, K. L. [375](#)
Preziosa, A. [225](#), [231](#)
Price, C. J. [123](#)
Price, J. [66](#), [67](#)
Price, J. H. [87](#), [347](#)
Price, R. H. [59](#)
Price, T. F. [229](#)
Priel, B. [176](#)
Priester, J. R. [33](#), [520](#)
Prinz, W. [323](#)
Prossin, A. R. [236](#)
Provost, J. [205](#)
Pryor, J. B. [316](#), [329](#), [330](#), [333](#)
Przybeck, T. R. [176](#)
Puff, C. R. [330](#)
Pugh, R. H. [499](#)
Purcell, C. [240](#)
Purvis, K. L. [688](#)

Qin, P. [150](#)
Quarton, R. J. [62](#)
Quené, H. [577](#)
Quera, V. [270](#), [355](#), [356–357](#), [359](#), [360](#), [363](#)
Quigley, K. S. [221](#), [230](#), [235](#), [236](#), [239](#), [240](#)

Rachmiel, T. B. [383](#)
Radford-Davenport, J. [384](#)
Raemaekers, M. [147](#)
Rafaeli, E. [86](#), [373](#), [376](#), [387](#), [390](#)
Raftery, A. E. [170](#)
Rahn, W. M. [408](#)
Raichle, M. E. [126](#)
Raines, B. E. [58](#), [60](#)
Rainie, L. [446](#)
Rajaratnam, N. [360](#), [473](#), [483](#)
Rakover, S. S. [15](#), [16](#)
Ramachandran, T. [231](#)
Ramamurthy, K. [206](#)
Ramdani, S. [275](#)
Rameson, L. T. [134](#)
Ramirez, M. D. [408](#)
Ramírez-Esparza, N. [347](#), [375](#)
Ramos, J. M. [408](#)
Ramsey, N. F. [147](#)
Rand, D. [532](#)
Ranganath, K. A. [464](#)
Rangel, A. [128](#)
Rankin, W. L. [423](#)
Rapkin, B. [73](#)
Rapoport, A. [191](#)
Rasch, G. [554](#)
Rasinski, K. A. [426](#)
Rathje, W. L. [86](#), [87](#)
Rauch, S. L. [227](#)
Raudenbush, S. W. [57](#), [162](#), [366](#), [392](#), [524](#), [525](#), [575](#), [592](#), [598](#)
Rawson, E. [149](#)
Ray, W. J. [235](#)
Raymond, P. [316](#), [329](#), [332](#), [333](#)
Rayner, K. [320](#)
Read, S. J. [254](#)
Reckase, M. D. [553](#)
Redcay, E. [127](#), [146](#)
Redelmeier, D. A. [378](#), [381](#)
Redner, R. [57](#), [74](#)

Reed, J. G. [683](#)
Reeder, L. G. [409](#)
Reeve, B. B. [560](#)
Reich, J. W. [67](#), [82](#)
Reichardt, C. S. [51](#), [53](#), [64](#), [69](#), [73](#)
Reicken, H. W. [196](#)
Reid, J. B. [347–348](#), [353](#)
Reid, K. [390](#)
Reidler, J. S. [231](#)
Reinecke, L. [445](#)
Reinsel, G. C. [67](#)
Reips, U.-D. [449](#), [458](#)
Reis, D. J. [238](#)
Reis, H. T. [28](#), [46](#), [74](#), [83](#), [85](#), [86](#), [87](#), [194](#), [195](#), [255](#), [374](#), [375](#), [376](#), [377](#), [380](#),
[382](#), [383](#), [384](#), [385](#), [387](#), [388](#), [389](#), [391](#), [445](#), [524](#), [589](#)
Reise, S. P. [484](#), [499](#), [518–519](#), [548](#), [566](#)
Reiss, D. [168–169](#)
Remmers, H. H. [422–423](#)
Ren, Y. [201](#)
Rendell, D. [228](#)
Reno, R. R. [57](#)
Rensvold, R. B. [547](#), [579](#), [581](#)
Rentfrow, P. J. [464](#)
Repetti, R. L. [347](#), [375](#), [379](#)
Restle, F. [198](#)
Reuter-Lorenz, P. A. [226](#)
Revilla, M. [424](#)
Reynolds, K. D. [72](#)
Reynolds, M. L. [516](#), [517](#)
Reynolds, S. [381](#)
Reynolds, T. J. [423](#)
Rezmovic, E. L. [489](#)
Rezmovic, V. [489](#)
Rhee, E. [419–420](#)
Rhemtulla, M. [539](#)
Rhodes, G. [241](#)
Rhodes, R. E. [423](#)
Rholes, W. S. [314](#), [319](#), [322](#)
Rhoton, Pat [409](#)

Ribisl, K. M. [53](#)
Rice, M. E. [473](#)
Rich, S. [159](#)
Richard, F. D. [690](#)
Richards, J. A. [269](#), [275](#)
Richards, M. H. [377](#), [384](#)
Richardson, D. C. [269](#), [271](#), [275](#), [276](#)
Richardson, J. D. [423](#)
Richardson, M. J. [254](#), [255](#), [256](#), [262](#), [264](#), [269](#), [271](#), [273](#), [274](#), [276](#)
Richardson, S. [170](#)
Richeson, J. [351](#)
Riddle, E. M. [102](#)
Riecken, H. [346](#)
Rieger, J. W. [151](#)
Ries, B. J. [223](#)
Riggs, S. A. [597](#)
Rijkes, C. M. [485](#)
Riley, M. A. [272](#), [273](#), [274](#)
Rilling, J. K. [146](#)
Rinaldi, S. [256](#)
Riolo, R. L. [267](#)
Riordan, M. A. [270](#)
Rips, L. J. [426](#)
Risch, N. [176](#)
Risheson, J. A. [3](#)
Ristikari, T. [680](#), [687](#)
Ritter, J. M. [382](#)
Riva, G. [225](#), [231](#)
Rizopoulos, D. [560](#)
Rizzolatti, G. [323](#)
Robbins, M. L. [347](#), [383](#), [387](#)
Robbins, T. W. [124](#)
Roberto, K. A. [415](#)
Roberts, I. [449](#), [454](#), [456](#), [457–458](#), [461](#), [464](#)
Roberts, J. K. [362](#), [592](#)
Robertson, C. [330](#)
Robertson, D. M. [226](#)
Robin, L. [378](#)
Robins, R. W. [373](#), [408](#), [425](#), [480](#), [482](#), [484](#), [485](#), [489–490](#), [534](#)

Robinson, B. F. [356](#), [360](#), [364](#)
Robinson, D. [431](#)
Robinson, J. P. [375](#), [383](#), [385](#)
Robinson, M. D. [225](#), [232](#)
Robinson, W. S. [572–573](#)
Robitzsch, A. [586](#), [587](#)
Rodgers, J. R. [238](#)
Rodrigues, A. [445](#)
Roenker, D. L. [330](#)
Roethlisberger, F. J. [194](#), [195](#)
Rogers, H. J. [484](#)
Rogers, P. [229](#)
Rogers, S. [419](#)
Rogers, T. B. [333](#)
Rogers, W. T. [507](#)
Rogge, R. D. [444](#), [445](#), [446](#), [453](#), [456](#), [457](#), [459](#), [463](#)

Rohde, S. [431](#)
Rohrbaugh, J. [210](#)
Roisman, G. I. [375](#), [383](#)
Roland, E. J. [207](#)
Rolffs, J. [446](#)
Rolls, E. T. [135](#), [241](#)
Roman, R. J. [334](#), [419–420](#)
Rook, D. [208](#)
Roos, L. L. [71](#)
Roos, N. P. [71](#)
Roscoe, J. [376](#)
Rose, N. [60](#)
Rosen, L. D. [445](#)
Rosenbaum, P. R. [56](#), [68](#), [69](#), [70](#), [71](#), [73](#)
Rosenberg, M. [480](#)
Rosenberg, M. J. [505](#)
Rosenblood, L. K. [375](#)
Rosenbloom, P. S. [315](#)
Rosenburg, M. J. [17](#)
Rosenkranz, M. A. [143](#)
Rosenthal, R. [17](#), [33](#), [44](#), [318](#), [350](#), [663](#), [667](#), [677–678](#), [681](#), [689](#), [691](#), [693](#), [696](#)
Roskos-Ewoldsen, D. R. [316](#), [332](#)
Rosnow, R. L. [33](#), [44](#), [667](#), [691](#)
Ross, C. E. [425](#)
Ross, D. C. [236](#)
Ross, L. [333](#), [696](#)
Ross, M. [378](#), [380](#), [381](#), [427](#)
Ross, M. W. [414](#)
Rosseel, Y. [586](#)
Rossi, J. S. [443](#)
Rossion, B. [141](#)
Roth, P. L. [688](#)
Rothbart, M. [326](#), [330](#)
Rotheram-Borus, M. [90](#)
Rothgeb, J. M. [428](#)
Rothschild, Z. K. [445](#)
Rothstein, H. R. [681](#), [682](#), [683](#), [693](#), [694](#), [698](#)
Rotshtein, P. [229](#)

Rotton, J. [681](#)
Rourke, D. L. [210](#)
Rovick, A. [204](#)
Rovine, M. J. [379](#), [577](#)
Rowe, G. [210](#)
Rowley, G. L. [483](#), [484](#)
Roy, A. K. [362](#), [395](#)
Rozelle, R. M. [58](#), [60](#)
Rozin, P. [87](#), [198](#)
Roznowski, M. [514](#)
Ruback, R. B. [207](#), [589](#)
Rubin, D. B. [49](#), [53](#), [55](#), [56](#), [64](#), [69](#), [88](#), [627](#), [628](#), [632–633](#), [638](#), [639](#), [678](#), [696](#)
Rubin, H. [378](#)
Rubin, S. [603–604](#)
Ruby, P. [231](#)
Ruch, G. M. [425](#)
Rucker, D. D. [517](#), [520](#), [524](#), [581](#), [655](#), [656](#), [671](#), [674](#)
Rudolph, T. J. [408](#)
Rugg, M. D. [138](#)
Ruhs, D. [463](#)
Ruiz-Belda, M. A. [238](#)
Rumelhart, D. E. [267](#)
Runkle, P. [193](#), [194](#), [197](#)
Rusbult, C. E. [2](#), [83](#), [85](#), [87](#), [515](#)
Rusby, J. [353](#)
Rushe, R. [365](#)
Russell, J. A. [115](#), [220](#), [221](#), [225](#), [227](#), [237](#), [238](#), [243](#)
Russell, M. [571](#), [574–575](#), [577](#)
Russell, R. L. [680](#)
Rutt, D. J. [384](#)
Ruud, P. A. [425](#)
Ryan, C. R. [615](#), [620](#), [623](#)
Ryan, C. S. [522](#), [665](#), [668](#), [669](#)
Ryan, K. [365](#)
Ryan, R. M. [376](#)
Ryff, C. D. [143](#)
Ryu, E. [57](#)

Saavedra, M. C. [445](#), [453](#)

Sabb, F. W. [148](#)
Sackman, H. [211](#)
Sadato, N. [124](#)
Sadikaj, G. [384](#)
Sadler, P. [268](#), [272](#)
Safarti, Y. [134](#)
Sagarin, B. J. [54](#)
Sailer, L. [381](#)
Saito, D. N. [124](#)
Salamon, G. [132](#)
Salas, E. [41](#), [207](#)
Salat, D. H. [132](#)
Salganik, M. J. [421](#)
Salih, F. A. [478](#)
Salmivalli, C. [586](#)
Salomon, K. A. [113](#)
Salovey, P. [515](#), [516](#)
Salt, P. [237](#)
Saltz, J. L. [118](#)
Saltzberg, J. A. [383](#)
Samaras, A. [206](#)
Samejima, F. [556](#)
Sampson, E. E. [20](#)
San Miguel, M. [255](#)
Sanbonmatsu, D. M. [316](#), [332](#), [333](#)
Sanchez, X. [222](#)
Sánchez-Meca J. [692](#), [693](#)
Sanders, A. F. [314](#)
Sanders, L. M. [410–411](#)
Sanderson, P. M. [203](#)
Sandoval, A. [210](#)
Sandy, C. J. [444](#)
Sandys, M. [200](#)
Sanfey, A. G. [146](#)
Sanford, R. N. [424](#)
Sangster, R. L. [418](#), [432](#)
Sansone, C. [444](#)
Santa, J. L. [39](#)
Santana, M. V. [269](#), [271](#), [273](#), [274](#), [275](#)

Sapsford, R. [406](#)
Sareen, J. [145](#)
Sargis, E. G. [444](#), [451](#)
Saris, W. E. [415](#), [422](#), [424](#), [431](#), [514](#)
Saron, C. D. [241](#)
Sassenberg, K. [663](#)
Satin, M. S. [376](#), [394](#)
Satorra, A. [514](#), [541](#)
Sauer, C. [147](#)
Sauter, D. [222](#), [227](#)
Savalei, V. [541](#)
Savoy, R. L. [126](#)
Saxbe, D. [146](#), [147](#), [375](#)
Saxe, L. [87](#), [101](#)
Saxe, R. [146](#)
Sayer, A. [631](#)
Sayer, A. G. [552](#)
Sayette, M. A. [204](#)
Sbarra, D. A. [376](#)
Scabini, D. [128](#)
Scahefer, A. [222](#)
Schaal, B. [324](#), [328](#), [335](#), [337](#)
Schachter, J. [224](#), [229](#), [230](#)
Schachter, S. [189](#), [190](#), [191](#), [196](#), [202](#)
Schacter, D. L. [137](#), [314](#)
Schacter, S. [346](#)
Schaefer, H. S. [127](#), [146](#)
Schaeffer, N. C. [419](#), [431](#)
Schafer, J. L. [69](#), [630](#), [632–633](#), [638](#), [647](#)
Schaffer, S. [4](#)
Schank, M. J. [383](#)
Scheef, L. [147](#)
Scheibe, S. [226](#)
Schemann, M. [238](#)
Scherer, K. R. [220](#), [221](#), [222](#), [226](#), [227](#), [228](#), [237](#), [242](#)
Scherg, M. [137](#)
Scherpenzeel, A. C. [445](#), [447](#), [462](#)
Scheufele, D. A. [413](#)
Schiffman, S. S. [516](#), [517](#)

Schilbach, L. [127](#)
Schillewaert, N. [424](#)
Schilling E. A. [376](#), [379](#)
Schimmack, U. [375](#), [384](#)
Schkade, D. A. [383](#)
Schloerschedit, A. M. [138](#)
Schlosser, A. [255](#)
Schmaling, K. B. [348](#)
Schmelkin, L. P. [480](#)
Schmid, C. H. [693](#)
Schmidt, F. L. [487](#), [682](#), [688](#), [689](#), [698](#)
Schmidt, J. A. [373](#)
Schmidt, R. C. [254](#), [255](#), [262](#), [263](#), [264](#), [269](#), [270](#), [273](#), [276](#)
Schmitt, B. M. [241](#)
Schmitt, N. [477](#), [479](#), [482](#), [659](#)
Schmittner, J. [375](#)
Schmukle, S. C. [599](#), [602](#)
Schneider, E. [224](#)
Schneider, M. [693](#)
Schneider, W. [205](#), [315](#)
Schneirla, T. C. [118](#)
Schoemann, A. M. [539](#)
Schomer, D. L. [240](#)
Schönbrodt, F. D. [602](#)
Schooler, J. [681](#)
Schopler, J. [376](#)
Schott, M. [581](#), [662](#)
Schreiber, C. A. [378](#), [381](#)
Schreiber, T. [271](#), [272](#)
Schroeder, D. A. [345](#)
Schuessler, K. F. [423](#)
Schuldberg, D. [254](#)
Schumacker, R. E. [512](#), [525](#), [527](#), [696](#)
Schuman, H. [381](#), [408](#), [410](#), [422–423](#), [425](#), [426](#), [431](#)
Schumann, D. [520](#)
Schurtz, D. R. [124](#), [690](#)
Schutter, D. J. [230](#)
Schütz, A. [445](#)
Schvaneveldt, R. W. [331](#), [332](#)

Schwalm, D. E. [62](#), [64](#)
Schwart, A. J. [147](#)
Schwartz, D. [602](#)
Schwartz, G. E. [116](#), [237](#), [241](#)
Schwartz, J. [115](#)
Schwartz, J. E. [381](#), [387](#), [389](#)
Schwartz, J. L. K. [445](#), [452](#)
Schwartz, R. D. [83](#), [86](#), [87](#), [395](#), [490](#)
Schwarz, G. [170](#)
Schwarz, J. C. [604](#)
Schwarz, N. [59](#), [224](#), [230](#), [335](#), [374](#), [378](#), [381](#), [383](#), [411](#), [422](#), [423](#), [425](#), [426](#)
Schwarzer, R. [491](#)
Scoboria, A. [684](#)
Scott, J. [201](#), [202](#), [425](#)
Scott, J. P. [159](#), [415](#)
Scott-Sheldon, L. A. [681](#)
Scratchley, L. S. [379](#)
Seal, D. W. [208](#)
Seamon, J. G. [321](#)
Sears, D. O. [20](#), [59](#), [83](#), [198](#), [444](#)
Seaver, W. B. [62](#)
Sechrest, L. [57](#), [74](#), [83](#), [86](#), [87](#), [395](#), [490](#)
Sedlmeier, P. [443](#)
Seeman, E. [164](#)
Seery, M. D. [111](#), [114](#)
Segal, N. L. [159](#)
Segal, S. J. [314](#)
Segall, M. H. [498](#)
Selig, J. P. [580](#)
Seligman, M. E. P. [444](#)
Semin, G. R. [256](#), [446–447](#)
Senchak, M. [382](#)
Senghas, A. [128](#), [224](#), [230](#)
Senulis, J. A. [241](#)
Seppa, C. [144](#), [149](#)
Seu, J. [204](#)
Sevincer, A. T. [419](#)
Sexter, M. [224](#)
Shadish, W. R. [11](#), [22](#), [27](#), [28](#), [36](#), [49](#), [50](#), [59](#), [60–61](#), [62](#), [64](#), [69](#), [73](#), [74](#), [88](#), [89](#),

90, 93, 395, 680, 681
Shaeffer, E. 424
Shaeffer, N. C. 419
Shafir, E. B. 149
Shah, D. V. 408
Shahar, G. 539
Shaklee, H. 191
Shallice, T. 331, 336
Sham, P. C. 178
Shamdasani, P. N. 208
Shanahan, J. 413
Shane, G. S. 209
Shannon, L. 381
Shapiro, D. 102
Shapiro, S. S. 611
Sharek, D. J. 463
Shattuck, D. W. 132
Shavelson, R. J. 481, 483, 484
Shaver, P. R. 2, 115, 385
Shavitt, S. 419, 420
Shaw, C. D. 258
Shaw, L. L. 85
Shaw, M. L. 39
Shea, B. J. 697
Shechter, D. 316
Sheets, V. 658–659, 662
Sheldon, K. M. 376
Shelley, K. S. 113
Shelling, T. 266
Shelton, J. N. 3, 351
Shen, F. 415
Sherif, C. 190, 191, 194
Sherif, M. 11, 191, 194, 420
Sherman, D. K. 445
Sherman, J. W. 373
Sherman, S. J. 114, 322, 323
Sherwin, R. G. 201
Sherwood, A. 112, 115
Sheu, C.-F. 560

Shevrin, H. [322](#)
Shiffman, S. [373](#), [379](#), [381](#), [382](#), [384](#), [389](#)
Shiffman, S. S. [378](#), [381](#), [384](#), [387](#), [388](#)
Shiffrin, R. M. [315](#)
Shinkareva, S. V. [151](#)
Shirako, A. [599](#)
Shiv, B. [318](#)
Shneiderman, B. [202](#)
Shockley, K. D. [269](#), [271](#), [273](#), [274](#), [275](#)
Shoda, Y. [254](#), [374](#), [380](#)
Shrout, P. E. [358–359](#), [379](#), [383](#), [387](#), [484](#), [523](#), [581](#), [656](#), [658](#), [659–660](#)
Shulka, R. K. [210](#)
Shulman, A. D. [84](#), [377](#)
Shultz, J. [86](#)
Sibley, C. G. [690](#)
Sidanius, J. [678](#)
Siegal, S. [191](#), [205](#)
Siegel, E. H. [221](#), [226](#), [227](#), [236](#), [240](#)
Siegel, J. T. [11](#)
Sigall, H. [101](#), [659](#), [663](#)
Sigman, M. [353](#)
Silver, B. [431](#)
Silver, B. D. [431](#)
Silver, J. S. [207](#)
Silver, R. D. [379](#)
Silverman, B. W. [63](#)
Silvestri, K. [66](#), [67](#)
Simmons, J. P. [443](#), [622](#)
Simmons, W. K. [144](#), [223](#), [228](#)
Simon, B. [463](#)
Simon, H. A. [322](#)
Simon-Thomas, E. R. [222](#), [227](#)
Simonsohn, U. [443](#), [622](#)
Simpser, A. [416](#), [421](#), [431](#)
Simpson, D. D. [330](#)
Simpson, J. A. [378](#)
Sinclair, R. C. [325](#)
Singer, B. H. [143](#)
Singer, E. [417](#), [418](#), [424](#), [427](#), [428](#), [429](#), [433](#)

Singer, J. D. [57](#), [453](#), [458](#), [582](#), [585](#)
Singer, J. E. [224](#), [229](#), [230](#)
Singer, M. [269](#)
Singh, B. H. [407](#)
Sinicropi-Yao, L. [222](#), [227](#)
Sinigaglia, C. [323](#)
Sitaram, R. [230](#)
Sivacek, J. [14](#)
Skelton, J. A. [325](#)
Skitka, L. J. [444](#), [451](#)
Skolnick Weisberg, D. [149](#)
Skowronski, J. J. [381](#)
Slade, R. [322](#)
Slagboom, P. E. [178](#)
Slatcher, R. B. [347](#), [375](#)
Slep, A. M. S. [365](#)
Sloan, D. M. [234](#)
Sloane, D. [364](#)
Sloboda, J. A. [228](#)
Slotter, E. B. [379](#), [383](#)
Slovic, P. [381](#)
Small, D. M. [127](#)
Smith, A. [275](#)
Smith, C. M. [200](#)
Smith, D. N. [161](#)
Smith, D. R. [201](#)
Smith, E. E. [149](#)
Smith, E. R. [39](#), [40](#), [194](#), [201](#), [254](#), [256](#), [267](#), [327](#), [378](#)
Smith, G. D. [693](#)
Smith, G. T. [487](#)
Smith, J. L. [446](#)
Smith, J. Z. [591](#)
Smith, K. J. [563](#)
Smith, L. B. [256](#), [277](#)
Smith, M. [221](#)
Smith, M. A. [202](#)
Smith, M. L. [681](#), [698](#)
Smith, N. K. [139](#)
Smith, P. B. [682](#)

Smith, P. L. [592](#), [598](#)
Smith, R. C. [223](#)
Smith, R. H. [346](#)
Smith, S. M. [131](#), [136](#), [445](#), [524](#)
Smith, T. L. [474–475](#), [489](#)
Smith, T. W. [417](#), [426](#)
Smith, V. K. [425](#)
Smolen, A. [162](#)
Smollan, D. [457](#)
Smyth, F. L. [464](#)
Smyth, J. D. [422–423](#), [430](#), [446](#)
Sneyd, J. [255](#)
Sniderman, P. M. [411](#)
Snidman, N. [493](#)
Snieszek, J. A. [211](#)
Snijders, T. A. B. [57](#), [575](#), [586](#), [592](#)
Snyder, C. R. R. [315](#), [316](#), [331](#), [336](#)
Snyder, C. W., Jr. [516](#), [517](#)
Snyder, D. J. [243](#)
Snyder, J. J. [347](#), [348](#), [365](#)
Snyder, M. [380](#), [487](#), [494](#), [590](#)
Sobel, M. E. [56](#), [547](#), [658](#)
Sobel, N. [225](#)
Soderstrom, E. J. [53](#)
Soellner, R. [73](#)
Sokol-Hessner, P. [147](#)
Solomon, B. [382](#)
Solomon, K. O. [229](#)
Solso, R. L. [315](#), [316](#), [331](#), [336](#)
Somerville, L. H. [148](#)
Sommer, B. [681](#)
Sommers, R. [191](#)
Sondheimer, R. M. [89](#)
Soneji, D. [230](#)
Song, A. W. [123](#)
Sörbom, D. [479](#), [542](#)
Sorrell, J. T. [223](#)
Sorrentino, R. M. [312](#)
Soto, C. J. [445](#), [464](#)

Spaulding, W. [323](#)
Spearman, C. [534](#), [535](#), [538](#), [565](#)
Spence, E. L. [118](#)
Spencer, M. E. [240](#)
Spencer, S. J. [44](#), [505](#), [663](#)
Sperling, G. [320](#)
Sperling, M. B. [604](#)
Spiegel, M. [586](#)
Spiegel, N. H. [229](#)
Spiegelhalter, D. [560](#)
Spitznagel, E. L. [357](#)
Spivey, M. J. [254](#), [274](#), [276](#)
Spotts, E. L. [169](#)
Sprecher, S. [378](#)
Spreng, R. N. [147](#), [228](#)
Sprott, J. C. [256](#)
Spruyt, A. [228](#)
Spunt, R. P. [146](#)
Srivastava, S. [443–444](#), [445](#), [446](#), [482](#), [495](#), [497](#), [504](#)
Srull, T. K. [267](#), [317](#), [322](#), [323](#), [330](#), [331](#)
St. Louis, R. D. [210](#)
Staats, M. [597](#)
Stadler, G. [394](#)
Stahl, D. [518](#)
Stam, C. J. [269](#)
Stanley, J. [109](#)
Stanley, J. C. [12](#), [22](#), [27](#), [40](#), [49](#), [61](#), [680](#)
Stapp, J. [414](#)
Stark, H. A. [137](#)
Stasser, G. [189](#), [200](#)
Steeh, C. [408](#)
Steel, P. [684](#)
Steel, R. P. [209](#)
Steele, C. M. [110](#), [350](#)
Steer, R. A. [242](#)
Steffen, V. J. [683](#)
Steiger, J. H. [170](#), [541](#)
Stein, M. B. [145](#)
Steinberg, L. [559](#), [560](#), [561](#), [563](#)

Steiner, D. D. [604](#)
Steiner, L. [192](#), [211](#)
Steiner, P. M. [70](#), [73](#), [74](#)
Steller, B. [324](#)
Stember, C. [431](#)
Stephan, F. F. [202](#)
Stephen, D. G. [275](#)
Stephenson, B. Y. [210](#)
Stepper, S. [224](#), [229](#)
Sterba, S. K. [539](#)
Stern, H. [147](#)
Stern, L. [224](#)
Stern, R. M. [235](#)
Sternberg, R. J. [491](#), [534](#), [697](#)
Sterne, J. A. [164](#)
Steve, K. W. [432](#)
Stevens, J. P. [527](#)
Stevens, S. S. [474](#)
Stevenson, R. A. [223](#), [236](#)
Stewart, D. W. [208](#)
Stewart, J. [90](#)
Stewart, L. [687–689](#), [695–696](#)
Stewart, T. L. [327](#)
Steyer, R. [60](#)
Stice, E. [127](#)
Stieger, S. [449](#), [452](#)
Stiff, C. [679](#)
Stigler, S. M. [678](#)
Stiller, J. [194](#)
Stillman, T. F. [124](#)
Stillwell, A. M. [329](#)
Stine, R. [658](#)
Stocks, E. L. [84](#), [85](#)
Stone, A. A. [373](#), [375](#), [376](#), [378](#), [379](#), [380](#), [381](#), [383](#), [384](#), [387](#), [388](#), [389](#), [390](#),
[394](#)
Stone, C. [563](#)
Stoolmiller, M. [365](#)
Storms, L. H. [314](#)
Storreston, M. [204](#)

Strack, F. [224](#), [229](#), [316](#), [329](#), [333](#), [423](#)
Strack, R. [322](#), [335](#)
Strack, S. [535](#)
Straf, M. L. [687](#)
Strahan, E. J. [505](#), [508](#), [509](#), [510](#)
Strauss, M. E. [537](#)
Streicher, V. J. [60](#)
Strejc, H. [384](#)
Strodtbeck, F. L. [189](#)
Stroebe, W. [193](#), [204](#), [207](#), [211](#)
Strogatz, S. H. [265](#), [277](#)
Strohmetz, D. B. [325](#)
Stroop, J. R. [335](#)
Stuart, E. A. [49](#), [59](#)
Su, Y-H [560](#)
Suci, G. J. [228](#)
Sudman, S. [381](#), [412](#), [422](#), [425](#), [426](#), [429](#), [431](#)
Suen, H. K. [360](#)
Sugawara, H. M. [515](#)
Suh, E. [379](#)
Sullivan, D. [445](#)
Sullivan, P. [605](#)
Suls, J. [373](#), [379](#), [383](#), [384](#)
Summers, K. J. [351](#)
Sumpter, D. J. T. [255](#)
Sun, C. R. [375](#), [379](#), [383](#)
Sun, X. [563](#)
Sundberg, J. [227](#), [237](#)
Suppes, P. [516](#), [518](#)
Sutton, A. J. [693](#)
Svrakic, D. M. [176](#)
Swaminathan, H. [484](#)
Swanenburg, K. L. [198](#)
Swann, W. B. [110](#)
Swanson, C. [254](#), [262](#), [365](#)
Swanson, K. [254](#), [258](#), [262](#), [365](#)
Swart, M. [225](#)
Swendeman, D. [90](#)
Swiedler, T. C. [204](#)

Swim, J. K. [379](#)
Swinth, K. R. [231](#)
Symons, C. S. [689](#)
Szapocznik, J. [73](#)

Tabachnick, B. G. [458](#), [459](#)
Tabibnia, G. [226](#)
Tahk, A. [412](#)
Tajfel, H. [11](#), [191](#)
Takagi, E. [200](#)
Takane, Y. [516](#), [517](#)
Takezawa, M. [201](#)
Talairach, J. [132](#)
Tamir, M. [228](#)
Tan, E. [221](#)
Tannenbaum, P. H. [228](#)
Tanner [318](#)
Tapp, J. [203](#)
Taraban, R. [210](#)
Taras, V. [684](#)
Tassinary, L. G. [102](#), [103](#), [104](#), [106](#), [119](#), [233](#), [234](#), [235](#), [239](#), [240](#), [269](#), [322](#)
Tateneni, K. [507](#)
Taylor, A. B. [658–659](#)
Taylor, B. J. [631](#), [632](#)
Taylor, D. A. [350](#)
Taylor, J. R. [423](#)
Taylor, S. E. [2](#), [147](#), [337](#), [378](#), [381](#), [384](#), [489](#)
Teachman, B. A. [225](#)
Tedlie, J. C. [494](#)
Tegie-Mocigemba, S. [228](#)
Telford, L. [348](#)
Tellegen, A. [159](#), [497](#), [545](#)
Telzer, E. H. [137](#)
Tennen, H. [373](#), [377](#), [384](#), [390](#), [453](#)
Terrin, N. [693](#)
Tesser, A. [254](#), [262](#), [265](#)
Tetlock, P. E. [197](#), [411](#)
Thein, R. D. [326](#), [327](#)
Theiner, G. [255](#)

Thelen, E. [256](#), [277](#)
Theraulaz, G. [255](#), [269–270](#)
Thiele, O. [451](#)
Thissen, D. [559](#), [560](#), [561](#), [563](#)
Thoemmes, F. J. [49](#), [57](#), [60](#), [73](#), [74](#)
Thomas, A. [560](#)
Thomas, D. L. [381](#)
Thomasson, N. [274](#)
Thompson, C. P. [330](#), [381](#)
Thompson, K. N. [696](#)
Thompson, R. L. [424](#), [458](#), [461](#)
Thompson, S. G. [627](#), [692](#), [694](#)
Thompson-Schill, S. L. [229](#)
Thornberry, T. [348](#)
Thorndike, R. M. [488](#)
Thurstone, L. L. [534](#), [535](#), [537](#), [538](#)
Thurstone, T. G. [537](#)
Tice, D. M. [124](#), [329](#)
Tickle-Degnen, L. [350](#)
Tidwell, M. C. O. [385](#)
Tieman, D. [330](#)
Tierney, J. [687–689](#), [695–696](#)
Tillman, R. [151](#)
Tindale, R. S. [200](#), [210](#)
Tingley, D. [60](#)
Tinsley, D. J. [491](#)
Tinsley, H. E. [491](#)
Tisak, J. [552](#)
Titus, L. J. [678](#)
Titus, W. [189](#)
Tiwan, H. [176](#)
Todd, M. [60](#)
Todd, P. [63](#)
Todorov, A. [2](#)
Tognoli, E. [269](#)
Tom, S. M. [134](#), [225](#), [227](#)
Toma, C. [671–672](#)
Tomaka, J. [112](#), [113](#), [114](#), [229](#), [240](#)
Tomarken, A. J. [143](#)

Tomlinson, J. M. [271](#), [276](#)
Tomlinson, M. [90](#)
Tompson, T. [412](#)
Tooke, W. [377](#)
Topmkins, P. [225](#)
Tormala, Z. L. [581](#), [655](#), [656](#)
Torre, K. [275](#)
Tortora, R. D. [431](#)
Tota, M. E. [34](#), [316](#), [320–321](#), [337](#)
Toth, J. P. [336](#)
Totterdell, P. [381](#)
Tourangeau, R. [425](#), [426](#), [429](#), [430](#), [433](#), [455](#)
Tournoux, P. [132](#)
Tracy, J. L. [373](#)
Traub, R. D. [137](#)
Traugott, M. W. [410](#), [418](#)
Trautwein, U. [573](#), [586](#), [587](#)
Treisman, A. M. [315–316](#)
Trepe, S. [445](#)
Trepel, C. [134](#)
Triandis, H. [498](#)
Trientes, R. J. H. [203](#)
Trochim, W. M. K. [64](#)
Troetschel, R. [323](#)
Troland, L. T. [104](#)
Trope, T. [326](#)
Tropp, L. R. [376](#), [383](#), [682](#)
Trull, T. J. [395](#)
Truman, D. B. [421](#)
Trzesniewski, K. H. [88](#), [406](#), [408](#), [425](#)
Tsai, A. [90](#)
Tsai, F. F. [445](#)
Tshibanda, L. [151](#)
Tsujioka, B. [539](#)
Tsutakawa, R. K. [560](#)
Tucker, C. [418](#)
Tucker, L. R. [507](#), [510](#), [541](#)
Tuerlinckx, F. [225](#), [375](#), [379](#)
Tukey, J. W. [159](#), [620–621](#), [669](#)

Tukey, P. A. [610](#)
Tulving, E. [137](#)
Tuma, A. H. [497](#)
Tumer, R. [128](#), [130](#)
Turco, R. [692](#)
Turk, C. L. [223](#)
Turkhan, J. S. [380](#)
Turkheimer, E. [159](#), [160](#), [162](#), [164](#), [175](#), [177](#), [178](#), [180](#)
Turner, C. F. [430](#)
Turner, J. [147](#)
Turner, T. J. [220](#)
Turrisi, R. [665](#)
Turvey, M. T. [263](#), [273](#), [274](#), [275](#), [276](#), [320](#)
Tversky, A. [381](#), [475](#)
Tweedie, R. [693](#)
Twenge, J. M. [690](#)
Tyler, R. B. [321](#), [332](#)
Tyson, R. [258](#), [262](#), [365](#)
Tzelgov, J. [656](#)
Tziner, A. [415](#)

Uchino, B. N. [505](#), [527](#), [529](#)
Uddin, L. Q. [134–135](#)
Uleman, J. S. [322](#), [334](#), [419–420](#)
Umansky, R. [176](#)
Umstad, M. P. [164](#)
Underwood, B. [23](#)
United Nations Economic and Social Commission for Asia and the Pacific [429](#)
United Nations Statistics Division [684](#)
Updegraff, J. A. [378](#), [384](#)
Urry, H. L. [143](#)
Uskul, A. K. [419](#)

Vaio, F. [259](#)
Valacich, J. S. [207](#)
Valdesolo, P. [224](#), [235](#)
Valentine, J. C. [678](#), [680](#), [682](#), [689](#), [697–698](#)
Vallacher, R. R. [254](#), [257](#), [259](#), [262](#), [263](#), [265](#), [266](#), [267](#), [276](#), [277](#)
Van Baaren, R. B. [350](#)

van Buuren, S. [632–633](#), [638](#)
Van de Ven, A. H. [209](#), [210](#)
Van de Vijver, F. [498](#)
Van De Walle, S. [426](#)
van den Bergh [577](#)
Van der Klaauw, W. [63](#)
van der Kouwe, A. [147](#)
van der Linden, W. J. [553](#)
Van der Maas, H. L. J. [265](#), [365](#)
van der Plight, J. [265](#)
Van der Vegt, G. S. [599](#), [602](#)
van Doomen, L. J. [112](#), [115](#)
van Engen, M. L. [684](#)
Van Essen, D. C. [144](#)
van Geert, P. [256](#)
Van Gelder, T. [276](#)
Van Heihnsbergen, C. C. [241](#)
Van Hoewyk, J. [417](#)
van Honk, J. [230](#)
Van Horn, J. D. [147](#)
Van Kempen, H. [409](#)
van Knippenberg, A. [4](#)
van Lange, P. A. M. [83](#), [85](#), [87](#)
Van Lighten, O. [224](#)
Van Meek, L. [681](#)
Van Orden, G. C. [256](#), [272](#), [275](#), [276](#)
Van Ryzin, G. G. [426](#)
van Veen, V. [142](#)
Van Vugt, M. [683](#)
van Wezel, R. J. A. [147](#)
Vanhaudenhuysen, A. [151](#)
VanHook, E. [335](#)
Vanman, E. J. [102](#), [103](#), [107](#), [117](#), [118](#), [120](#), [234](#)
Vanzetti, N. A. [364](#)
Vaollone, R. [84](#)
Varnell, S. P. [57](#)
Vaslow, J. B. [689](#)
Vaughan, J. [205](#)
Vazdarjanova, A. [238](#)

Vazire, S. [347](#), [375](#), [443–444](#), [445](#), [446](#), [604](#)
Veit, R. [230](#)
Veldhuijzen, W. [605](#)
Velicer, W. F. [65](#), [67](#), [535](#)
Velleman, P. F. [609](#), [619](#)
Velten, E. [223](#)
Vêncio, R. Z. N. [147](#)
Verette, J. [2](#)
Vevea, J. L. [684](#), [692](#), [693](#), [694](#)
Vexler, D. [89](#)
Vicary, A. M. [375](#), [383](#)
Viechtbauer, W. [698](#)
Viger, S. G. [226](#)
Villani, D. [225](#), [231](#)
Vink, M. [147](#)
Vinkyuzen, A. A. E. [178](#)
Vinokur, A. D. [59](#)
Visscher, P. M. [177](#), [178](#)
Visser, P. S. [418](#)
Viswanathan, A. [128](#)
Vivian, D. [353](#)
Vogeley, K. [127](#)
Vohs, K. D. [146](#), [384](#), [447](#)
Vollrath, D. A. [200](#)
von Bracken, H. [375](#)
von Davier, M. [539](#)
von Eye, A. [541](#)
Von Eye, V. [383](#)
Voogt, R. J. J. [409](#)
Voracek, M. [449](#), [452](#)
Vowles, K. E. [223](#)
Vrana, S. R. [118](#)
Vranceanu, A. M. [380](#), [384](#)
Vuilleumier, P. [241](#)
Vul, E. [144](#)
Vytal, K. [241](#)

Wachter, K. W. [687](#)
Wager, T. D. [126](#), [144](#), [146](#), [147](#), [227](#), [228](#), [232](#), [241](#), [242](#)

Waggener, T. B. [269](#)
Wagner, D. D. [148](#)
Waid, W. M. [102](#)
Waksberg, J. [431](#)
Walden, T. [203](#)
Waldron, M. [164](#), [180](#)
Waldstein, S. R. [111](#)
Wallace, J. M. [183](#)
Wallach, M. A. [30](#)
Waller, N. G. [419](#), [420](#), [484](#), [498](#)
Wallin, B. G. [239](#)
Wallot, S. [256](#), [275](#)
Walsh, V. [230](#)
Walster, E. [189](#)
Walster, G. W. [189](#), [210](#)
Walter, F. [599](#), [602](#)
Walther, J. B. [447](#)
Walton, G. M. [89](#)
Walton, M. A. [53](#)
Walton, R. [177](#)
Wamboldt, F. [362](#)
Wampold, B. E. [348](#), [362](#), [363](#)
Wan, C. K. [34](#), [383](#)
Wang, C. [176](#)
Wang, M. C. [688](#)
Wang, N. [415](#)
Wang, R. [416](#), [421](#), [431](#)
Wang, S. W. [347](#)
Wang, W.-C. [560](#)
Ware, J. E., Jr. [424](#)
Ware, J. H. [395](#)
Warlaumont, A. S. [269](#), [275](#)
Warner, R. M. [269](#), [272](#), [364](#), [365](#)
Warner, W. G. [424](#)
Warneryd, B. [424](#)
Warren, J. A. [118](#)
Warren, W. H. [254](#), [276](#)
Wasel, W. [324](#), [328](#), [335](#), [337](#)
Wason, P. C. [424](#)

Wasseman, E. [230](#)
Wasserman, S. [202](#), [605](#)
Watanabe, Y. [200](#)
Waters, E. B. [360](#)
Watkins, S. C. [409](#)
Watson, C. B. [577](#)
Watson, D. [237](#), [375](#), [377](#), [379](#), [383](#), [409](#), [537](#), [635](#), [682](#)
Watts, D. J. [87](#)
Waugh, C. E. [251](#)
Way, B. M. [225](#), [227](#)
Waytz, A. [125](#)
Webb, E. J. [83](#), [86](#), [87](#), [395](#), [490](#)
Webb, N. M. [483](#), [484](#)
Webber, C. L. [274](#)
Weber, A. L. [188](#)
Weber, H. [321](#)
Weber, J. [148](#)
Weber, K. [537](#)
Webster, D. M. [420–421](#)
Webster, G. D. [124](#), [589](#)
Wegener, D. T. [33](#), [322](#), [425](#), [505](#), [506](#), [507](#), [508](#), [509](#), [510](#), [520](#), [524](#), [527](#), [529](#),
[530](#), [535](#), [538](#), [560](#)
Wegner, C. [147](#)
Wegner, D. M. [314](#), [328](#), [336](#)
Weick, K. E. [194](#)
Weijden, T. [605](#)
Weijters, B. [424](#)
Weinberger, J. [381](#)
Weingart, L. [198](#)
Weingart, L. R. [198](#), [592](#), [598](#)
Weinstein, S. [230](#), [240](#)
Weisberg, H. F. [406](#), [408](#)
Weisbuch, M. [111](#), [114](#)
Weiskopf, N. [128](#), [230](#)
Weiss, A. [221](#), [237](#)
Weiss, R. L. [351](#), [353](#), [360](#)
Welch, F. [409](#)
Weldon, E. [198](#)
Wellens, T. R. [325](#)

Wells, A. [53](#)
Wells, G. A. [697](#)
Welsch, R. E. [619](#)
Wendorf, C. A. [577](#)
Weng, L.-J. [424](#)
Wenger, M. J. [272](#)
Wentland, E. J. [381](#)
Wentz, R. [449](#), [454](#), [455](#), [456](#), [457–458](#), [461](#), [464](#)
Wenze, S. J. [385](#)
Wenzel, G. [23](#)
Wesman, A. G. [425](#)
Wessel, I. [205](#)
West, B. J. [275](#)
West, S. G. [49](#), [51](#), [54](#), [56](#), [57](#), [59](#), [60](#), [62](#), [64](#), [67](#), [70](#), [72](#), [73](#), [74](#), [75](#), [84](#), [90](#), [395](#),
[482](#), [491](#), [494](#), [513](#), [521](#), [522](#), [523](#), [548](#), [646](#), [658–659](#), [662](#), [665](#), [667](#), [668](#),
[669](#), [690](#)
West, T. V. [598](#), [602](#)
Westenberg, P. M. [485](#)
Westfall, J. [602](#)
Whalen, P. J. [148](#), [227](#)
Whaley, S. E. [353](#)
Wharton, J. L. [65](#), [67](#)
Wheatley, T. [124](#)
Wheelan, S. A. [188](#)
Wheeler, L. [54](#), [375](#), [382](#), [384](#)
Wheeler, R. E. [143](#)
Wherry, R. J. [539](#)
Whisman, M. A. [666](#), [669](#)
White, B. [191](#), [194](#)
White, D. [227](#)
White, H. D. [682](#)
White, I. R. [627](#)
White, J. [365](#)
White, R. K. [190](#)
Whitehurst, G. J. [359](#)
Whitley, B. R. [685](#)
Whybrow, P. C. [254](#)
Whyte, W. F. [194](#)
Wichman, A. L. [582](#), [585](#)

Wicker, A. W. [678](#)
Wicklund, R. A. [325](#)
Widaman, K. F. [491](#), [499](#), [506](#), [508](#), [509](#), [510](#), [528](#), [534](#), [535](#), [538](#), [539](#), [544](#), [548](#),
[552](#)
Wiebe, E. N. [463](#)
Wieder, G. B. [360](#)
Wig, G. S. [148](#)
Wiggins, J. S. [487](#), [489](#), [491](#)
Wiitala, W. L. [690](#)
Wikman, A. [424](#)
Wilbarger, J. L. [235](#)
Wilcox, J. B. [526](#)
Wilcox, K. J. [159](#)
Wilcox, R. R. [609](#)
Wilde, M. [586](#)
Wiles, J. [275](#)
Wiley, John [518](#)
Wilhelm, F. H. [269](#), [364](#), [387](#)
Wilk, M. B. [610](#), [611](#)
Wilke, H. A. M. [204](#)
Wilkinson, L. [633](#)
Willemssen, A. [203](#)
Willet, J. B. [57](#)
Willett, J. B. [453](#), [458](#), [582](#), [585](#)
Williams, C. J. [332](#)
Williams, J. [466](#), [658](#)
Williams, K. J. [383](#)
Williams, K. D. [146](#), [190](#), [452](#)
Williams, L. A. [235](#)
Williams, L. E. [4](#), [146](#)
Williams, L. M. [226](#), [230](#), [235](#)
Williams, M. L. [207](#)
Williams, S. C. R. [147](#), [226](#)
Williams, S. L. [480](#), [485](#)
Williford, A. [586](#)
Willis, G. B. [428](#)
Wilms, M. [127](#)
Wilson, D. [23](#)
Wilson, D. B. [681](#), [682](#), [683](#), [689](#), [690](#), [691](#), [698](#)

Wilson, D. C. [426](#)
Wilson, D. T. [560](#)
Wilson, M. [365](#)
Wilson, M. F. [225](#)
Wilson, T. D. [21](#), [60](#), [83](#), [151](#), [317](#), [322](#), [427](#), [445](#)
Wilson-Mendenhall, C. D. [223](#), [228](#)
Windle, M. [548](#)
Winer, B. J. [27](#), [334](#)
Winkielman, P. [144](#), [235](#)
Winkler, J. D. [424](#)
Winship, C. [69](#)
Winter, J. P. [207](#)
Wirth, R. J. [539](#), [553](#), [560](#)
Wiseman, F. [430](#)
Wish, M. [515](#), [516–517](#), [518](#)
Wittenbaum, G. M. [211](#)
Wittenbrink, B. [602](#)
Wohlschlager, A. M. [230](#)
Woike, B. A. [383](#)
Wojcieszak, M. E. [425](#)
Wolfe, C. [205](#)
Wolff, H. G. [102](#)
Wolford, G. [40](#), [322](#)
Wolford, G. L. [147](#)
Wolfram, S. [266](#)
Wong S. P. [484](#)
Wong, E. C. [147](#)
Wong, J. J. [54](#)
Wong, M. M. [385](#)
Wong, S. [60–61](#), [73](#)
Wong, S. P. [359](#), [696](#)
Wong, V. C. [69](#), [73](#), [74](#)
Wood, A. M. [627](#)
Wood, C. [679](#)
Wood, J. V. [383](#)
Wood, R. [560](#)
Wood, R. E. [376](#)
Wood, W. [688](#), [698](#)
Woodman, T. [225](#)

Woodruff, D. [478](#)
Woods, C. M. [456](#), [563](#)
Woodward, C. K. [57](#), [68](#)
Woodward, J. A. [57](#), [68](#)
Woody, E. [268](#), [272](#)
Woodzicka, J. A. [83](#)
Woolrich, M. W. [131](#)
Word, C. O. [663](#)
World Bank [684](#)
World Values Survey [684](#)
Worsley, K. [136](#)
Wortman, C. B. [379](#)
Wray, N. R. [178](#)
Wright, B. D. [560](#)
Wright, B. R. E. [161](#)
Wright, G. [210](#)
Wright, M. F. [445](#)
Wright, S. D. [413](#)
Wyer M. M. [379](#), [383](#), [384](#)
Wyer, R. S., Jr. [267](#), [317](#), [322](#), [323](#), [330](#), [331](#)

Xavier Castellanos, F. [134–135](#)
Xiao, R. [177](#)
Xu, J. [128](#)

Yalch, R. F. [409](#)
Yalcin, B. [177](#)
Yamagishi, M. [424](#)
Yamagishi, T. [200](#)
Yamamoto, M. [365](#)
Yammarino, F. J. [598](#), [602](#)
Yan, Z. [454](#), [457](#), [458](#), [461](#), [464](#)
Yang, H. C. [207](#)
Yang, J. [178](#)
Yang, K. [498](#)
Yang, S. [202](#)
Yaniv, I. [211](#)
Yarkoni, T. [144](#)
Ye, C. [417](#)

Ye, F. [147](#)
Yeager, D. S. [416](#), [431](#)
Yeaton, W. [57](#), [74](#)
Yee, P. L. [205](#)
Yen, W. M. [561](#)
Yeshurun, Y. [225](#)
Yik, M. S. M. [243](#)
Yip, T. [376](#)
Yokum, S. [127](#)
Yon, R. [413](#)
Yoo, S.-S. [230](#)
Yoshimoto, D. [365](#)
Young, A. W. [226](#)
Young, B. [409](#)
Young, F. W. [516](#), [517](#)
Yuan, K. [513](#)
Yun-Tein, J. [539](#)
Yunokuchi, K. [240](#)
Yzerbyt, V. Y. [520](#), [524](#), [660](#), [662](#), [671–672](#), [673–674](#)

Zabel, Jeffery [409](#)
Zagorsky, Jay [409](#)
Zajonc, R. B. [2](#), [36](#), [190](#), [326](#)
Zalcman, S. S. [236](#)
Zandbelt, B. [147](#)
Zanna, M. P. [19](#), [44](#), [326](#), [351](#), [382](#), [422–423](#), [505](#), [663](#)
Zareba, W. [384](#)
Zbilut, J. P. [274](#)
Zedeck, S. [360](#)
Zeger, S. L. [362](#)
Zelaya, F. O. [147](#)
Zelazo, P. D. [124](#), [137](#)
Zellinger, P. M. [424](#)
Zembrodt, I. M. [515](#)
Zemke, P. E. [204](#)
Zhang, G. [507](#)
Zhang, S. [509](#), [510](#), [538](#), [671](#)
Zhang, X. [226](#)
Zhang, Z. [365](#), [393–394](#), [525](#), [587](#)

Zhong, X. [513](#)
Zhou, Z. H. [56](#)
Ziglio, E. [210](#)
Ziliak, James P. [409](#)
Zilles, K. [144](#)
Zillman, D. [17](#)
Zimbardo, P. G. [86](#), [348](#)
Zimmerman, P. H. [203](#)
Zimmermann, M. G. [255](#)
Zimowski, M. F. [560](#)
Zlochower, A. J. [367](#)
Zochowski, R. [259](#)
Zubieta, J.-K. [236](#)
Zuckerman, A. [383](#)
Zuckerman, M. [229](#)
Zurowskis, B. [230](#)
Zyphur, M. J. [393–394](#), [525](#), [587](#)

Subject Index

- action interference program (AIP) [291–292](#)
- Actor-Partner Interdependence Model (APIM) [595–597](#)
 - example [597](#)
 - extended to groups [598–599](#)
 - mediation and moderation [597–598](#)
 - patterns [597](#), [598](#)
- Affective Misattribution Procedure (AMP) [288–289](#)
- associative-propositional evaluation (APE) model [296](#)

- Beck Depression Inventory [242](#)
- behavior genetic research methods alternative parameterization, genetically informed phenotype regression [172–173](#)
- heritability [175](#), [179](#)
- interpretation, standardization [173–175](#)
- matched pairs [161](#)
- methodological era [159–160](#)
- modeling sequence [170–172](#)
- personality as nonexperimental science [160–161](#)
- random effects models [162–165](#)
- religiosity, delinquency in MZ twins [161–162](#)
- research conduct recommendations [179–180](#)
- research methods, overview [159](#)
- structural equation models, ACE regression [166–170](#)
- behavior genetic research methods, molecular genetic approaches candidate gene studies [176–177](#)
 - genome-wide association studies (GWAS) [177–178](#)
 - genome-wide complex trait analysis (GCTA) [178](#)
 - linkage analysis [175–176](#)
- behavioral observation and coding acceptable reliability, required time [360–361](#)
 - advances, future research directions [367](#)
 - analog observation [348–349](#)
 - coding systems, units [351–352](#)
 - data analysis [361–362](#)
 - defined [345](#)

- dynamical systems modeling [365, 366](#)
- experimental manipulation [349–351](#)
- Finns *r*, [359](#)
- Interpersonal Process Code (IPC) example [353](#)
- intraclass correlation (ICC) [358–359](#)
- live observation [345](#)
- molar vs. molecular approach [352–353](#)
- multilevel loglinear analysis [366–367](#)
- multilevel survival analysis [365–366](#)
- multiple dimensions [352–353](#)
- naturalistic observation [346–347](#)
- ordinal, interval and ratio observations, summary and recommendations [360](#)
- quasi-naturalistic observation [346–347](#)
- reliability across observations, contexts, and time [360](#)
- research method value, reasons for [345–346](#)
- training observers [354](#)
- validity [361](#)
- weighted kappa [360](#)
- behavioral observation and coding, interrater agreement [354–355](#)
 - ACI statistic [358](#)
 - categorical observations [355–358](#)
 - choice [355](#)
 - Cohen's kappa [355](#)
 - Holley and Guilford's *G*, [357](#)
 - for sequences [360](#)
 - Van Eerdewegh's *V*, [357](#)
 - weighted kappa [356–357](#)
- behavioral observation and coding, sequential analysis [362–363](#)
 - behavior sequences dimensional analysis [364–365](#)
 - bidirectional dependence [362](#)
 - data structure [362–363](#)
 - dominance [362–363](#)
 - loglinear approach [363–364](#)
 - time domain approach [364–365](#)
 - unidirectional dependence [362](#)
- Cattell-Horn-Carroll (CHC) model [534](#)
- causal inference, Rubin's Causal Model (RCM) [49–57](#)
 - analysis by treatment [55](#)

- assumptions, problems and remedies 53–57
- causal effect estimation, ideal case and randomized experiment (RE) a_1 51
- group administration of treatment 57
- intention to treat analysis (ITT) 54–55
- local average treatment (LATE) 55–56
- participants measured at posttest, no attrition 53–54
- participants received full treatment as assigned, Treatment Adherence 54–56
- randomization as approach to fundamental problem of causal inference 50–51
- randomization, illustrative example 51–53
- randomization, properly carried out 53
- Stable-Unit-Treatment-Value Assumption (SUTVA) 56–57
- treatment concompliance effects 55–56
- causal processes, mediational analysis 18–19
- causal relationships, generalization strategies 57–61
 - causal explanation 60
 - discrimination validity 59–60
 - empirical interpolation, extrapolation 60
 - extra-statistical approaches, five principles 59–60
 - generalization, formal statistical model 58
 - heterogeneous irrelevancies 59
 - proximal similarity 59
 - statistical strategy, sampling from defined population 58–59
- complex dynamical systems complexity theory 253
 - researchers, advances 254
 - traditional approach 253
- computer assisted qualitative data analysis software (CAQDAS) 203
- confirmatory factor analysis (CFA) 565–566
 - data model 536
 - data requirements for 537–539
 - exploratory factor analysis (EFA) and 535–536
 - mean, covariance structure model 536–537
 - measured variables selection 514
 - model evaluation 513–514
 - model modification 514
 - model specification 512
 - observations selection 537–538
 - origins, history 536
 - research participants selection 514–515

- variables selection, domain representation, scales, items, parcels [538–539](#)
- confirmatory factor analysis (CFA), empirical examples multitrait-multimethod (MTMM) matrix [543–547](#)
 - personality data [542–543](#)
- confirmatory factor analysis (CFA), factorial invariance additional invariance forms [551–552](#)
 - configural invariance [548](#)
 - empirical example [549–551](#)
 - strict factorial invariance [548](#)
 - strong factorial invariance [548](#)
 - theoretical requirements [547–549](#)
 - weak factorial invariance [548](#)
- confirmatory factor analysis (CFA), implementation evaluation [541–542](#)
 - latent variables (LV), manifest variables (MV) [536](#), [538–539](#)
 - maximum likelihood (ML) estimation [540–541](#)
 - readjustment [542](#)
 - specification [539–540](#)
- construct validity, conceptual replications [17–18](#)
- construct validity, construct to/from operation causal processes, mediational analysis [18–19](#)
 - construct validity, conceptual replications [17–18](#)
 - multiple operations, convergent and discriminant validity [18](#)
- dyads and groups data design and analysis distinguishability [591–592](#)
 - dyadic reciprocity [605](#)
 - independence of observations [590](#)
 - one-with-many (OWM) design [603](#)
 - phenomenon-assimilation agreement [589](#)
 - variables types [590–591](#)
- dyads and groups data design and analysis, Actor-Partner Interdependence Model (APIM) [595–597](#)
- dyads and groups data design and analysis, Actor-Partner Interdependence Model (APIM), example [597](#)
- dyads and groups data design and analysis, Actor-Partner Interdependence Model (APIM), mediation and moderation [597–598](#)
- dyads and groups data design and analysis, Actor-Partner Interdependence Model (APIM), patterns [597](#), [598](#)
- dyads and groups data design and analysis, group studies actor effect [600](#)
 - dyadic outcomes [603](#)

- dyadic outcomes, social relations model (SRM) [599–602](#)
 - partner effect [600–601](#)
 - relationship effect [601](#)
 - single-measure outcomes [603](#)
 - two-measure outcomes [603](#)
- dyads and groups data design and analysis, multilevel modeling (MLM) [592](#)
 - dyadic data [594–595](#)
 - group data [592–594](#)
 - negative nonindependence [595](#)
- dynamical systems analysis autocorrelation, cross-correlation [272–273](#)
 - behavioral measurement [268–270](#)
 - cross-recurrence analysis [273–275](#)
 - fractal analysis [275–276](#)
 - further reading [276–277](#)
 - lagged correlation [272–273](#)
 - linear, nonlinear times-series analysis [272](#)
 - methods [270–276](#)
 - qualitative and graphical assessment [270–272](#)
 - recurrence analysis [273–275](#)
 - relative phase analysis [273](#)
 - spectral analysis, cross-spectral coherence [272](#)
- dynamical systems modeling [365](#), [366](#)
 - attractors [262–263](#)
 - bifurcations [264–266](#)
 - cellular automata, agent-based and artificial neural network models [266–268](#)
 - difference equations [258–259](#)
 - differential equations [259–262](#)
 - order and control parameters [263–264](#)
- dynamical systems theory chaos [257–258](#)
 - interaction-dominant dynamics [256](#)
 - nonlinearity [256–257](#)
 - self-organization [255](#)
 - soft-assembly [255–256](#)
- electroencephalogram (EEG)/event related potentials (ERP) data analysis [142–143](#)
 - data averaging/cleaning/preprocessing [141–142](#)
 - data collection [140–141](#)
 - EEG replication [147](#)

- study design [137–140](#)
- electronic brainstorming (EBS) [207](#)
- emotion and affect measurement Affective Norms for English Words (ANEW) [228](#)
 - Beck Depression Inventory [242](#)
 - best practices tips, tricks and secrets [243–244](#)
 - emotion, affect definitions [220](#)
 - emotion as mental state [220](#)
 - International Affective Digitized Sounds (IADS) [227](#)
 - Project EMMA (Engaging Media for Mental Health Applications) [230](#)
 - Trier Social Stress Test (TSST) [230](#)
- emotion and affect measurement, evoked states measurement [231–232](#)
 - autonomic nervous system activity [238–240](#)
 - behavior [238](#)
 - central nervous system activity [240–241](#)
 - endocrine, immune, inflammatory changes [241](#)
 - facial muscle activity, facial EMG [234–237](#)
 - observer ratings [237–238](#)
 - subjective experiences [241–243](#)
 - vocal acoustics [237](#)
- emotion and affect measurement, methods for inducing changes bodily movements and posture [229](#)
 - confederates [230](#)
 - faces [226–227](#)
 - films [221–225](#)
 - imagery and recall [228](#)
 - images [226](#)
 - motivated performance tasks [230–231](#)
 - music [228](#)
 - physiological manipulations [229–230](#)
 - real world stimuli [231](#)
 - virtual reality [231](#)
 - words [228–229](#)
- event-contingent recording [384](#)
- everyday experience in natural context, studying methods administration format [387–388](#)
 - aggregations, composites [391–392](#)
 - checklists [388](#)
 - complex multilevel models [393–394](#)

- compliance documentation [389](#)
- computerized devices [387–388](#)
- conceptual rationale [374–377](#)
- conceptualizing everyday experience [377–380](#)
- data analytic strategies, considerations [390–391](#)
- diary data as self-reports [382](#)
- event-contingent recording [384](#)
- everyday experience methods [373](#)
- examples [375](#)
- hypothesis tests [375–376](#)
- instrument design [388–389](#)
- interval-contingent recording [382–383](#)
- methodological rationale [374](#), [380–382](#)
- multilevel modeling logic [392–393](#)
- multiple operationalism and [395](#)
- participant issues [389–390](#)
- pragmatic considerations [385–386](#)
- programmatic research, methods integration with [395–396](#)
- protocol design [386–387](#)
- protocols comparison [384–385](#)
- protocols types [382](#)
- rating scales [388](#)
- recency, salience, sense-making, state of mind [381](#)
- record-keeping impact on experience [390](#)
- research aims [374–377](#)
- signal-contingent recording [383–384](#)
- technology impact, techniques [373–374](#)
- temporal pattern analysis [394–395](#)
- theory complexity, future directions [396–397](#)
- exploratory factor analysis (EFA) [505–506](#), [535–536](#)
 - confirmatory analysis (CFA) vs., [510–512](#)
 - statistical issues [506–509](#)
- external validity ecological validity, representativeness of [21](#)
 - external validity importance [22](#)
 - relevance, significance of [21–22](#)
 - robustness replication [19–21](#)
- extrinsic affective Simon task (EAST) [289–290](#)
- field experiment methods advantages vs. disadvantages [91](#)

- designs addressing field challenges 89–90
- downstream field experimentation 89
- encouragement designs 88
- hybrid lab-field experiments 89
- randomization, control in field settings 87–88
- randomized rollout designs 88–89
- field research advantages 84–85
 - causal testing and 85
 - constructs definition 84
 - defined 82
 - inductive power and 84–85
 - laboratory research vs., 82–85
 - relevance and 85
 - theoretical maps, development and 81
 - theory's pragmatic worth testing 85
- field research, observational methods ethnography and 86
 - individual characteristics 86–87
 - individual, population characteristics estimates 86–87
 - interviews, key informants and 86
 - participant observation 86
 - personal observation 85–86
 - population characteristics 87
 - qualitative methods 85–86
 - situation characteristics observations 87
- field research, practical issues Institutional Review Board (IRB) 94
 - memorandum of understanding 94
 - stakeholders position 91–94
- field research, quasi-experimentation in field interrupted times series analysis 90
 - regression discontinuity 90
- fMRI methods Brodman's areas (BAs) 136–137
 - conjunction analysis 134
 - coregistration 131
 - data acquisition 128
 - data analysis 133–135
 - data cleaning and preprocessing 129–133
 - DICOM files 130
 - familywise error rate (FWE) vs. false discovery rate (FDR) 135–136
 - fieldmap correction 130
 - fMRI replication 147

- functional/effective connectivity [134](#)
- normalization [132](#)
- parametric modulation [134](#)
- power and sample size [126–127](#)
- raw data conversion [130](#)
- realignment [130–131](#)
- reorientation [132](#)
- second-level models [135](#)
- spatial smoothing [132–133](#)
- statistical thresholding [135–136](#)
- study design [125–128](#)
- technical limitations [127–128](#)

- genome-wide association studies (GWAS) [177–178](#)
- genome-wide complex trait analysis (GCTA) [178](#)
- go no/go association task (GNAT) [289](#)

- Implicit Association Test (IAT) [284–286](#)

 - evaluative priming task [286](#)
 - semantic [286](#)

- implicit measures absolute vs. relative interpretations [300–301](#)

 - action interference program (AIP) [291–292](#)
 - Affective Misattribution Procedure (AMP) [288–289](#)
 - approach-avoidance tasks [290–291](#)
 - associative-propositional evaluation (APE) model [296](#)
 - as behavior predicting tools [293–295](#)
 - behavior prediction mechanisms [303–304](#)
 - as bias in information processing prediction tool [293–295](#)
 - conscious vs. unconscious representations [297](#)
 - context effects [299](#)
 - convergence vs. divergence between [304–305](#)
 - experimental manipulations, automatic effects [300](#)
 - explicit, implicit measures dissociations [298](#)
 - extrinsic affective Simon task (EAST) [289–290](#)
 - as formation, change of mental representations tool [296](#)
 - future research directions [303–305](#)
 - go no/go association task (GNAT) [289](#)
 - Implicit Association Test (IAT) [284–286](#)
 - Implicit Association Test (IAT), evaluative priming task [286](#)

- Implicit Association Test (IAT), semantic [286](#)
- implicit measures, defined [283–284](#)
- implicit relational assessment procedure (IRAP) [291–292](#)
- measure procedure selection [293](#)
- multiple processes underlying [301–303](#)
- old vs. new representations [297–298](#)
- self-report measures [283](#)
- semantic priming tasks [288](#)
- social desirability, faking, lie detection [298–299](#)
- sorting paired features (SPF) task [291](#)
- individual difference scaling (INDSCAL) [516](#)
- intention to treat analysis (ITT) [54–55](#)
- International Affective Digitized Sounds (IADS) [227](#)
- Internet research [443](#), [451](#)
 - attention, compliance [446](#)
 - client-side programming [448–449](#), [450](#)
 - custom formatting [449–450](#)
 - data collection costs [444](#)
 - designs less Internet-amenable [446–447](#)
 - future research directions [466–467](#)
 - immediate feedback, study programming [450](#)
 - Internet hosting service cost [449](#)
 - Internet hosting service questions [449–451](#)
 - Internet-based study hosting services [449](#)
 - participants, answer change prevention [450](#)
 - participants, more than one sitting [450–451](#)
 - rare populations study groups [444–445](#)
 - research automation [444](#)
 - research challenges [445–447](#)
 - sample representativeness [445–446](#)
 - sample sizes, data collection benefits [443–445](#)
 - server- vs. client-side programming [448–449](#)
 - student samples vs., [443–444](#)
 - study data access [451](#)
 - study data security, storage [451](#)
 - study design software [449](#)
 - study designs support and [445](#)
 - study web pages creation [447–449](#)
 - study web pages hosting [447](#)

Internet research, online experiments implementation attention engagement 457
attention measurement 456–457
contact information collection 453
crowdsourcing, Mechanical Turk (MTurk) 462–463
data cleaning 458–459
data security 465
deception, debriefing 466
dropout management 458
dynamic/interactive manipulations 452
email distribution lists, listservs 461
email invitations, response rates enhancement 461
ethical issues 464
feedback as recruitment incentive 464
follow-up participation, briefer options 454
greater than minimal risk studies, 466
inattentive, invalid responses screening 459
indirect measures, tasks online 452
informed consent 464–465
Inquisit 3 Web 452
invitation automation 453–454
IP addresses, anonymity 465–466
longitudinal data, linking waves 453–454
longitudinal, experience sampling studies 453–454
mandatory questions 456
multiple contacts use for follow-up waves 454
multiple submissions screening 458–459
online advertising 461
online forums, websites 459–460
online studies, website lists 460–461
outliers screening 459
presentation formatting, diverse platforms 455–456
pretesting 458
probability-based Internet panels 462
professional presentation 456
public vs. private behavior 465
PxLab 453
questions ordering 455
questions per page 455
random assignment 451

- rapport building [454](#)
- recruitment, sampling strategies [459](#)
- snowball recruitment [463–464](#)
- study design considerations [454–455](#)
- study length [457–458](#)
- text manipulations [451–452](#)
- Interpersonal Process Code (IPC) [353](#)
- interval-contingent recording [382–383](#)
- intraclass correlation (ICC) [358–359](#)
- item response theory (IRT) [534](#), [565–566](#)
 - Classic Test Theory vs., [552](#)
 - data requirements [559](#)
 - empirical example [561–563](#)
 - item, test information [559](#)
 - model implementation [560–561](#)
- item response theory (IRT), dichotomous item response models item characteristic curves (ICCs) [553–556](#)
 - one-parameter logistic model (1PLM) [552–554](#)
 - three-parameter logistic model (3PLM) [555–556](#)
 - two-parameter logistic model (2PLM) [554–555](#)
- item response theory (IRT), measurement invariance empirical example [565](#)
 - general requirements, Differential Item Functioning (DIF) [563–565](#)
- item response theory (IRT), polytomous item response models Generalized
 - Partial Credit Model [558](#)
 - Graded Response Model (GRM) [556–558](#)
 - Modified Graded Response Model [558](#)
 - Normal Response Model (NRM) [559](#)
 - Operating Characteristic Curves (RCCs) [556–558](#)
 - Partial Credit Model (PCM) [558–559](#)
 - Rating Scale Model [558](#)
 - Response Characteristic Curves (RCCs) [556–558](#)
- laboratory research advantages of [82–83](#)
 - complex systems in laboratory and [84](#)
 - culture of laboratory and [84](#)
 - disadvantages of [83–84](#)
 - observations [83](#)
 - participant units [83](#)
 - settings [83–84](#)

- treatments [83](#)
- latent growth curves [582–583](#)
 - alternative parameterizations [584–585](#)
 - interpretation [583–584](#)
- local average treatment (LATE) [55–56](#)
- measurement general considerations [473–474](#)
 - model building and evaluation definition [474](#)
 - psychometric, representational approaches [474–475](#)
- measurement, construct validation consequential validity evidence [488–489](#)
 - content validity [487](#)
 - evidence types for [487–489](#)
 - external validation, convergent and divergent aspects [487](#)
 - generalizability evidence [488](#)
 - integrated conception of [485–487](#)
 - LOTS [490–491](#)
 - method variance [489–490](#)
 - multitrait multimethod matrix (MTMM) [489](#)
 - structural validity evidence [488](#)
 - substantive validity evidence [487–488](#)
 - validity, traditional definitions [485](#)
- measurement reliability, generalizability [476](#)
 - attenuation correction [482](#)
 - classic test theory [476](#)
 - coefficient alpha [477–481](#)
 - generalizability theory [483–484](#)
 - item characteristic curve [484](#)
 - item response theory (IRT) [484–485](#)
 - psychometric data reporting [482–483](#)
 - reliability evidence types [476–477](#)
- measurement, scale construction model testing questionnaire question construction issues [496–499](#)
 - SEM-based measurement model constructs, confirmatory factor analysis (CFA) [491–496](#)
- mediation and moderation definitions [653–654](#)
 - from measured to manipulated mediators [662–664](#)
 - mediated moderation [673–674](#)
 - mediation assumptions [656–658](#)
 - mediation basic analytic model [654–656](#)

- mediation indirect effects estimation, testing [658–659](#)
- moderated mediation [673–674](#)
- moderation definitions, basic models [664–666](#)
- moderation, interaction detection difficulty [669–671](#)
- moderation interpretation, presentation [664–666](#)
- moderation, multilevel interactive models [671–672](#)
- moderation, simple slopes testing [666–667](#)
- moderation, standardized slopes [669](#)
- multilevel mediation [659–660](#)
- observed vs. latent variable models [659–660](#)
- relation, distinction [653](#)
- testing process by interaction strategy (TPIS) [663–664](#)
- meta-analysis arithmetic means across studies standardization [690–691](#)
 - arithmetic means vs. standardized mean difference effect sizes [691](#)
 - biased methods, effect sizes correction [688–690](#)
 - coding reliability [684](#)
 - conceptual analysis of literature [679–680](#)
 - conducting, evaluating [697](#)
 - cultural, social structural characteristics [684](#)
 - effect size calculations reliability [688](#)
 - effect size indexes of association [684–685](#)
 - effect sizes gauging associations direction [685–687](#)
 - effect sizes in individual studies estimation [684](#)
 - future in social-personality psychology [698](#)
 - large research literatures review [682](#)
 - locating relevant studies [682–683](#)
 - multiple cultures studies [682](#)
 - multiple reports from individual studies [687](#)
 - narrative reviewing [674](#)
 - nonreported results [688](#)
 - quantitative review history [678](#)
 - quantitative synthesis process [679](#)
 - quantity's magnitude, arithmetic means use [690](#)
 - reported statistical information precision [687–688](#)
 - research synthesis, additional resources [697–698](#)
 - studies samples, boundaries setting [680–682](#)
 - study characteristics [683–684](#)
 - unpublished studies [681–682](#)
- meta-analysis, meta-analytic database effect size indexes of association

- interpretations [696–697](#)
- effect size magnitude depictions [694–695](#)
- mean effect size, homogeneity of effect sizes [691–693](#)
- meta-analytic moderators, models testing [694](#)
- nonindependent effect sizes [695–696](#)
- outlier diagnoses [694](#)
- preliminary considerations [691](#)
- publication bias evaluation [693–694](#)
- missing data analysis analysis phase [638–639](#)
 - auxillary variables [643–644](#)
 - deletion methods [633](#)
 - imputation phase [636–638](#)
 - imputation phase, variable selection [638](#)
 - imputed data sets, serial dependence [637–638](#)
 - incomplete cases, accuracy improvement [642–644](#)
 - maximum likelihood estimation [640–642](#)
 - mean imputation, available items averaging [633–635](#)
 - methodologists vs. researchers [627](#)
 - Missing at Random (MAR) mechanism [629–630](#)
 - Missing Completely at Random (MCAR) mechanism [628](#)
 - missing data mechanisms [628](#)
 - multiple imputation [636](#)
 - multiple imputation choice [645–646](#)
 - multiple imputation vs. maximum likelihood estimation [644–645](#)
 - Not Missing at Random (NMAR) mechanism [630](#)
 - Not Missing at Random (NMAR)-based analyses [646–648](#)
 - planned missing data designs [631](#)
 - planned missing data designs, repeated measures [632](#)
 - pooling phase [639–640](#)
 - practical considerations [645](#)
 - regression imputation [635](#)
 - standard practice [627](#)
 - stochastic regression imputation [635–636](#)
 - three-form design [631–632](#)
 - traditional missing data handling methods [632–633](#)
- multidimensional scaling (MDS) [515](#)
 - dimensionality determination [516–517](#)
 - model choice [515–516](#)
 - solutions evaluation, design issues [517–518](#)

- solutions evaluation, interpretation [517](#)
- solutions evaluation, proximity data collection [517–518](#)
- solutions evaluation, research participants selection [518](#)
- multilevel and longitudinal modeling aggregation, disaggregation [572–573](#)
 - atomistic fallacy [573](#)
 - contextual effects [573](#)
 - cross-sectional designs [577](#)
 - ecological fallacy [573](#)
 - intraclass correlation coefficient (ICC) [576](#)
 - levels [571–572](#)
 - longitudinal designs [577](#)
 - multilevel modeling (MLM) basics [573–575](#)
 - multilevel structural equation modeling (MSEM) [585–586](#)
 - nested data [571](#), [586–587](#)
 - nested data, common research designs [577](#)
 - observation independence and [571](#)
 - structural equation modeling (SEM) basics [573–574](#)
 - structural equation modeling (SEM) vs. multilevel modeling (MLM) [576](#), [585–586](#)
- multilevel and longitudinal modeling, panel designs [577–579](#)
 - configural invariance [579](#)
 - latent growth curves [582–583](#)
 - latent growth curves, alternative parameterizations [584–585](#)
 - latent growth curves, interpretation [583–584](#)
 - mediation testing [580–582](#)
 - predictive paths testing [580](#)
 - strong invariance [579–580](#), [582](#)
 - weak invariance [579](#)
- multilevel modeling (MLM) [524–525](#), [530](#)
- multiple operations, convergent and discriminant validity [18](#)
- multitrait-multimethod (MTMM) matrix [543–547](#)
- nasty, ill-mannered data example [623–625](#)
 - multiple groups [615](#)
 - multiple regression [619](#)
 - remedies, nonparametric statistics [620](#)
 - remedies, outliers [621–622](#)
 - remedies, transformations [620–621](#)
 - robustness, reason for concern [608–609](#)

- simple regression [615–616](#)
- simple regression, homogeneity of variance [618–619](#)
- simple regression, outliers [616–618](#)
- single groups, normality assumption [608](#)
- single groups, outliers [612–613](#)
- source of problems [608](#)
- statistical method abuse [625](#)
- two groups [613](#)
- two groups, homogeneity of variance [613–615](#)
- two groups, outliers detection [615](#)
- new perspectives representation group identities [3](#)
 - idea sources [3](#)
 - intellectual puzzles [1–2](#)
 - personal puzzles [2–3](#)
 - worldview defense [3](#)
- nominal group technique (NGT) [209–210](#)
- nonexperimental data, causal and noncausal hypotheses causal hypothesis,
 - summary of methods [529–530](#)
- causal hypothesis types [519–520](#)
- confirmatory factor analysis (CFA), measured variables selection [514](#)
- confirmatory factor analysis (CFA), model evaluation [513–514](#)
- confirmatory factor analysis (CFA), model modification [514](#)
- confirmatory factor analysis (CFA), model specification [512](#)
- confirmatory factor analysis (CFA), research participants selection [514–515](#)
- direct cause hypothesis [521–522](#)
- exploratory factor analysis (EFA) [505–506](#)
- exploratory factor analysis (EFA), statistical issues [506–509](#)
- exploratory factor analysis (EFA) vs. confirmatory analysis (CFA) [510–512](#)
- factor extraction [506–507](#)
- factor rotation [508–509](#)
- individual difference scaling (INDSCAL) [516](#)
- inferring causality conditions [520–521](#)
- measured variables selection [509–510](#)
- mediated moderation [520](#), [524](#)
- mediational relations [522–523](#)
- moderated mediation [519–520](#), [523–524](#)
- moderational relations [523](#)
- multidimensional scaling (MDS) [515](#)
- multidimensional scaling (MDS), dimensionality determination [516–517](#)

- multidimensional scaling (MDS), model choice [515–516](#)
- multidimensional scaling (MDS) solutions evaluation, design issues [517–518](#)
- multidimensional scaling (MDS) solutions evaluation, interpretation [517](#)
- multidimensional scaling (MDS) solutions evaluation, proximity data collection [517–518](#)
- multidimensional scaling (MDS) solutions evaluation, research participants selection [518](#)
- multilevel modeling (MLM) [524–525](#), [530](#)
- multiple linear regression [521](#)
- noncausal methods, summary and comparison [518–519](#)
- nonexperimental studies benefits [504–505](#)
- number of factors selection [507](#)
- parallel analysis [507–508](#)
- relative strength of direct causes [522](#)
- research participants selection [510](#)
- RMSEA model fit measure [514–515](#)
- scree test [507](#)
- statistical methods, experimental vs. nonexperimental [505](#)
- structural equation modeling (SEM) [525](#), [530](#)
- structural equation modeling (SEM), design issues [529](#)
- structural equation modeling (SEM), direct cause relations [527](#)
- structural equation modeling (SEM), hybrid hypotheses [529](#)
- structural equation modeling (SEM), mediation [527–528](#)
- structural equation modeling (SEM) model fitting, evaluation, modification [526–527](#)
- structural equation modeling (SEM), model specification [525–526](#)
- structural equation modeling (SEM), moderation [528–529](#)
- one-with-many (OWM) design [603](#)
- priming and automaticity research [311–312](#), [336–337](#)
 - demand characteristics and mindset priming [325–326](#)
 - goal and behavior priming, beyond perception [323–324](#)
 - perceptual experience, individual differences in [312–313](#)
 - perceptual experience, internal states influence [312](#)
 - priming, automaticity together [316](#)
 - priming manipulations strength [323](#)
 - priming research techniques [316–317](#)
 - priming research techniques, mindset priming [317](#), [324–325](#)

- priming topics [324](#)
- supraliminal vs. subliminal priming [322](#)
- unwanted effects of priming [325](#)
- priming and automaticity research, automaticity research roots goal-dependent automaticity, skill-acquisition research [314–315](#)
- preconscious processing [315–316](#)
- priming and automaticity research, automaticity research techniques control attempts and uncontrollability [336](#)
- efficiency [326](#)
- manipulation of attentional demands [327–329](#)
- measurement of efficiency [326–327](#)
- memory organization clustering measures [329–331](#)
- response latencies as dependent variable [333–335](#)
- sequential priming techniques [331–333](#)
- uncontrollability [333–335](#)
- unintended processing effects [329–333](#)
- priming and automaticity research, conceptual priming [316–317](#)
- masking [320–321](#)
- subliminal priming [319–322](#)
- supraliminal planning [317–319](#)
- priming and automaticity research, priming research roots priming in social psychology [314](#)
- recent experience as individual difference [313–314](#)
- Project EMMA (Engaging Media for Mental Health Applications) [230](#)
- psychophysiology affective states: positive and negative affect, constructs [115](#)
- affective states: positive and negative affect, context [115](#)
- affective states: positive and negative affect, one-to-one relationships [115–116](#)
- assessment and summary, accrued advantages [119](#)
- assessment and summary, best ideas [119](#)
- assessment and summary, gold standard and [119–120](#)
- assessment and summary, implementation [119](#)
- attitude functionality [114–115](#)
- basic physiological processes [106](#)
- catalog of valid indexes [120](#)
- cellular process [106–107](#)
- conditioning [108](#)
- control systems, neural process structure and function [106](#)
- design implications [110](#)
- endocrine process [107](#)

- epistemological issues [103](#)
- facial EMG indexes of positive and negative affect, rationale [116–117](#)
- facial EMG indexes of positive and negative affect, validation research [116](#)
- invariance [103–104](#)
- motivational states, challenge and threat [110](#)
- one-to-one relationship [105](#)
- other indexes [119](#)
- physiological responses [105](#)
- psychological constructs [105](#)
- rational, constructs [110](#)
- rational, context [110–111](#)
- rational, one-to-one relationships [110–111](#)
- recording [108–109](#)
- research examples, belief in just world [114](#)
- research examples, prejudice and discrimination, technology [116–118](#)
- research examples, stigma [114–115](#)
- specific methodological concerns [109](#)
- startle eye-blink reflex indexing, rationale [118](#)
- startle eye-blink reflex indexing, validation research [118](#)
- startle reflex responses, technology [118–119](#)
- technical background [107](#)
- technical background, physical response signals [107](#)
- technical background, sensors [108](#)
- technical background, signal path [107–108](#)
- technical background, transducers [108](#)
- validation research, correlational studies [112–113](#)
- validation research, experimental studies [113](#)
- validation research, manipulated physiology studies [113–114](#)
- validation research, predictive validation studies [114](#)
- validation research, summaries [114](#)
- validity threats, instrumentation [109–110](#)
- validity threats, testing [109](#)

- quasi-experimental designs
 - alternative non-randomized designs effectiveness [73–74](#)
 - assignment rule adherence verification [63](#)
 - Campbell's perspective, second approach to causal inference [61–62](#)
 - causal generalization [64, 73](#)
 - delayed effects [67](#)

- design elements, pattern matching 66–67
- equating groups strategies 68–70
- functional form specification 63–64
- GPA regression example 62–63
- history, selection, instrumentation 65
- internal validity threats 70–72
- internal validity threats, identification 65–66
- interrupted time series design (ITS) 64–65
- nonequivalent control group designs, observational studies 68
- regression discontinuity (RD) design 62–63, 64
- statistical power 64, 68, 72–73
- time series modeling, statistical issues 67

research

- choosing, framing, competition 4–5
- everyday interest, scientific advance 4
- evidence promotion importance 5
- puzzle solving satisfaction 4
- rules of evidence, discovery 3–4
- story telling, entertainment 5
- research design Campbellian tradition 44–45
 - data analysis and 27
 - experimental design 27
 - nonexperimental, passive observational design 27
 - quasi-experimental design 27
 - research unit 45–46
 - tactics vs. tactics 46
 - three fundamental categories 27–28
 - unequals Ns, 43
- research design, confounds and artifacts demand characteristics, experimenter bias 43
 - designs for mediation studying 44–45
 - differential treatment-related attrition 43–44
 - social comparisons among conditions 44
- research design, dependent variables 39–42
 - average across multiple items 40
 - items as factor 40–41
 - items selection for dependent measure 39–40
 - MANOVAs 41–42

- pretesting [39–40](#)
- separate item analysis [41](#)
- single item use [40](#)
- structural equation model [41](#)
- research design, extremity of levels [30–31](#)
 - naturally occurring events matching [31](#)
 - number of levels [31–32](#)
 - powerful manipulations [31](#)
- research design, independent variables (factors) [28–32](#)
 - factor treatment decision [29–30](#)
 - fixed vs. random factors [28–30](#)
 - fixed vs. random factors, definitions [28–30](#)
 - participants, other random factors [29](#)
 - random factors power [30](#)
- research design, power typical designs, low power [42](#)
 - ways to increase [43](#)
- research design, relations among factors [32–39](#)
 - analysis with covariates [35](#)
 - between-participants vs. within-participants use [32, 33](#)
 - conceptual replication [37](#)
 - counterbalancing [37–38](#)
 - covariates with power [34–35](#)
 - crossed, nested, confounded factor [32](#)
 - crossing factors reasons, error variance reduction (statistical conclusion validity) [34–35](#)
 - crossing factors reasons, generality of effect (external validity) establishment [35–37](#)
 - crossing factors reasons, testing theoretically predicted interactions [33–34](#)
 - data transformations, interpretations complication [34](#)
 - interactions vs. conditioned means interpretation [33](#)
 - Latin square designs [38–39](#)
 - nested factors [39](#)
 - observation nonindependence, within-participants designs [37](#)
 - power for interactions [33–34](#)
 - replication success estimations [37](#)
 - theoretically predicted dissections demonstration [33](#)
 - varying a factor, across studies [36–37](#)
 - varying a factor, systematically [36](#)
 - varying a factor, unsystematically [36](#)

research design, validity issues [11–13](#)

construct validity, construct to/from operation [15–19](#)

demonstration purposes [11](#)

event correlation [11](#)

independent vs. dependent variable [12](#)

internal validity, third variable problem [12](#)

theory, operation links [15](#)

theory-testing research cycle [15](#)

utilitarian perspective, cause-effect relationships [11–12](#)

research reasons [1, 6](#)

collaboration [5](#)

publish or perish [5](#)

research funding, grant writing [5–6](#)

service to humanity [6](#)

spiritual quests [6](#)

teaching [6](#)

Rubin's Causal Model (RCM) [49–57](#)

signal-contingent recording [383–384](#)

small research group methods audio-video hardware, software [203–204](#)

Bank Wiring Room study [195–196](#)

computer assisted qualitative data analysis software (CAQDAS) packages [203](#)

computer programmers [205](#)

computer simulation [200–201](#)

computer technology, data collection at arbitrary group tasks [204–206](#)

Delphi technique [210–211](#)

electronic brainstorming (EBS) [207](#)

focus groups [207–208](#)

general purpose software programs [205](#)

generic strategies, field and archival research [193–194](#)

generic, strategies, observational field methods [194–195](#)

group brainstorming [206–207](#)

group effectiveness [211](#)

group, group contexts [188, 192](#)

groups as context/means for research aid application [206–211](#)

innovative methods [201–202](#)

Judge-Advisor Systems approach [211](#)

nominal group technique (NGT) [209–210](#)

nonparticipant observation [195–196](#)

- participant observation [196](#)
- quality circles (QCs) [208–209](#)
- social network analysis [201–202](#)
- sociometry [201](#)
- software acquisition [204–205](#)
- structural properties of groups [201](#)
- surveys, interviews [199–200](#)
- task groups [192](#)
- Time by Event Member Pattern Observation (TEMPO) [195–196](#)
- small research group methods, archival studies experimental methods [197–198](#)
 - field experiments [196–197](#)
 - systematic observation of groups [198–199](#)
- social and affective neuroscience, conceptual issues brain as predictor:
 - correlation vs. prediction [149–150](#)
 - data variability [148–149](#)
 - experimental vs. ecological validity [146](#)
 - fMRI and EEG replication [147](#)
 - forward and reverse inference [143](#)
 - mind reading [150–151](#)
 - rhetorical power of neuroimaging data [149](#)
 - spuriously high correlations [144–146](#)
- social and affective neuroscience, questions can answer brain mapping [124](#)
 - convergence [124–125](#)
- social and affective neuroscience, research methods future themes, core issues [151–152](#)
 - history, introduction and overview [123–124](#)
- social and affective neuroscience, types of questions methods can answer
 - divergence [125](#)
- social relations model (SRM) [599–602](#)
- Stable-Unit-Treatment-Value Assumption (SUTVA) [56–57](#)
- structural equation modeling (SEM) [525](#), [530](#)
 - design issues [529](#)
 - direct cause relations [527](#)
 - hybrid hypotheses [529](#)
 - mediation [527–528](#)
 - model fitting, evaluation, modification [526–527](#)
 - model specification [525–526](#)
 - moderation [528–529](#)
- survey research coverage error [433](#)

- data from students [404–405](#)
- definition, methodology [404](#), [406](#)
- funding, empirical justification [405](#), [406](#)
- Internet, web and [405](#)
- laboratory studies and [405](#)
- measurement error [433](#)
- Mechanical Turk (MTurk) [405](#)
- nonresponse error [433](#)
- processing error [433](#)
- questionnaire design literature [433–434](#)
- questionnaires [404](#)
- sample population [405](#)
- sampling error [433](#)
- specification error [433](#)
- survey datasets [406](#)
- total survey error [432–433](#)
- survey research, data collection computer-assisted personal interviewing (CAPI) [429](#)
 - computer-assisted telephone interviewing (CATI) [429](#)
 - cost [430](#)
 - data quality [431](#)
 - desired response rate [430](#)
 - face-to-face interviews [429](#)
 - length of period [430](#)
 - population characteristics [430](#)
 - question content [430](#)
 - question form [430](#)
 - questionnaire length [430](#)
 - sampling strategy [430](#)
 - self-administered questionnaires [429–430](#)
 - staff, facilities availability [430–431](#)
- survey research, interviewing interviewer selection [431](#)
 - interviewer training [431–432](#)
 - supervision [431–432](#)
 - validation [432](#)
- survey research, pretesting behavior coding [427–428](#)
 - cognitive interviewing [428](#)
 - conventional pretesting [427](#)
 - interview-administered questionnaires, pretesting methods [427–428](#)

- pretesting methods compared [428](#)
- self-administered questionnaire pretesting [428–429](#)
- survey research, question design and measurement error no-opinion filters, attitude strength [425](#)
- open vs. closed questions [422–423](#)
- order of response alternatives [425](#)
- question order [426](#)
- question wording [425](#)
- questions to avoid [426–427](#)
- rating scale formats [424–425](#)
- rating vs. ranking [423](#)
- survey research, sampling cluster sampling [414–415](#)
 - coverage error [419](#)
 - data via Internet [415](#), [421](#)
 - haphazard sampling [420](#)
 - multistage sampling [420–421](#)
 - nonprobability sampling [412](#), [419–422](#)
 - nonresponse error [417–419](#)
 - probability sampling [411–412](#)
 - purposeful sampling [420–421](#)
 - quota sampling [421](#)
 - robo polls [419](#)
 - sampling error [416–417](#)
 - simple random sampling [413](#)
 - snowball sampling [421](#)
 - stratified sampling [414](#)
 - systematic sampling [413–414](#)
 - typical sampling methods [415–416](#)
- survey research, study designs Affect Misattribution Paradigm [408–409](#)
 - combined use, cross-sectional and panel surveys [409–410](#)
 - cross-sectional surveys [406–408](#)
 - experiments within surveys [410–411](#)
 - experiments within surveys, benefits [411](#)
 - implicit measurement [411–412](#)
 - mood, life satisfaction [411](#)
 - panel surveys [408–409](#)
 - projection hypothesis [408–409](#)
 - racism [410–411](#)
 - repeated cross-sectional surveys [408](#)

Time by Event Member Pattern Observation (TEMPO) [195–196](#)

Trier Social Stress Test (TSST) [230](#)

validity, optimizing types of isolation vs. construct validity [24](#)

lab vs. field setting [22–24](#)

multivariate, multilevel research [24–25](#)